

Summer 2015

A fuzzy logic approach to teacher performance measured by principal evaluations

Ashley Johnson Moran

Follow this and additional works at: http://csuepress.columbusstate.edu/theses_dissertations



Part of the [Educational Leadership Commons](#)

Recommended Citation

Moran, Ashley Johnson, "A fuzzy logic approach to teacher performance measured by principal evaluations" (2015). *Theses and Dissertations*. 212.

http://csuepress.columbusstate.edu/theses_dissertations/212

This Dissertation is brought to you for free and open access by CSU ePress. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of CSU ePress.

**A FUZZY LOGIC APPROACH TO TEACHER PERFORMANCE
MEASURED BY PRINCIPAL EVALUATIONS**

By
Ashley Johnson Moran

A Dissertation
Submitted in Partial Fulfillment
of the Requirements for
the Degree of Doctor of Education
in Curriculum and Instruction

Columbus State University
Columbus, GA

June 2015

ABSTRACT

While principals and current methods used to evaluate teacher performance have been found to accurately identify the teachers with students who have the largest and smallest achievement gains in their schools, they are less able to distinguish among teachers in between these extremes. A majority of states have implemented new reform in teacher evaluation strategies in order to diagnose and improve teacher performance, but unfortunately, have not incorporated a method capable of accurately representing qualitative factors and the relationships between them. The main objective of this study was to examine to what degree a fuzzy logic expert system could elicit and quantify teacher performance using state teacher evaluation methods.

Theoretical background is given as a foundation for fuzzy logic expert systems in order to understand their implications to teacher performance evaluation. The experimental results from a case study of one Georgia high school were compared with the traditional state evaluation method. The findings of the study indicate a fuzzy logic expert system may help identify teachers lying at the boundary of two different categories of performance and show potential for a fuzzy logic expert system to provide a more accurate continuum of teacher performance.

DEDICATION

To my wonderful mother and loving husband

When I started this journey, I would have never known what a blessing it is to still have them both in my life. It was their strength, faith, and determination through their own battles that inspired and motivated me.

ACKNOWLEDGEMENTS

I would like to thank the following members of my committee. Without their tireless efforts in providing me support and guidance, this study would not have been possible.

I would like to express the deepest appreciation to my committee co-chair Professor Iris Saltiel. If it were not for her encouragement in my pursuit to explore an area of research “outside the box” and her willingness to take on such an endeavor, this study would have never materialized from the start.

I would like to thank Dr. Deborah Gober, who undertook this great time commitment to co-chair my committee despite her many other academic and professional commitments. It was her invaluable comments and insights that were needed to improve the numerous drafts of this paper.

I would like to thank Dr. Shamim Khan, who introduced me to fuzzy logic and also mentored me through independent studies of this area of research. I am incredibly grateful for his hours of patience in answering my many questions.

I would also like to thank Dr. Camille Bryant, who, in addition to participating in the committee, has been my instructor for several courses. It was with great respect for her teaching that her questions stimulated me to examine areas of the research even further.

I could not have dreamed of a better committee. It is to each of them that I owe my deepest appreciation.

TABLE OF CONTENTS

List of Figures	viii
List of Tables	x
CHAPTER ONE-INTRODUCTION	
Introduction.....	1
Evaluation Reform in the United States.....	2
Georgia Teacher Evaluation Reform	3
Problem Statement.....	5
Principal Evaluations	5
Combining Measures of Performance	5
Rationale	7
Advantages of Fuzzy Logic in General	7
Advantages of Fuzzy Logic in Education	10
Purpose of Study	13
Research Questions	14
Conclusion	14
CHAPTER TWO-FUZZY LOGIC AND FUZZY SETS	
Introduction to Fuzzy Logic.....	16
Fuzzy Logic and Set Theory	16
Membership Functions.....	17
Fuzzy Rules.....	20
Design of a Fuzzy Logic Expert System.....	20

Step 1: Fuzzification of the Input Variables	22
Step 2: Assigning Rules	22
Step 3: Aggregating all Outputs.....	24
Step 4: Defuzzification	25
CHAPTER THREE-REVIEW OF LITERATURE	
Introduction.....	27
Teacher Performance Measures	27
Value-added Models	27
Classroom Observations	32
Principal Evaluations	34
Scoring Method.....	39
Weighting Criteria	40
Fuzzy Logic Expert Systems in Education	43
Review of Studies Using Fuzzy Logic Applied to Student Assessment	44
Review of Studies Using Fuzzy Logic Applied to Teacher Performance	52
Validity and Reliability of Fuzzy Logic Expert Systems	58
Conclusion	61
CHAPTER FOUR-METHODS	
Introduction.....	62
Research Questions	62
Participants.....	63
Instrumentation	64
Validity and Reliability of Teacher Assessment Performance Standards	66

Design of the Fuzzy Logic Expert System	70
Knowledge Acquisition Phase	71
Step 1: Fuzzification of the Input Variables	74
Step 2: Assigning Rules	76
Step 3 and 4: Aggregation and Defuzzification of the Output Variables	81
Testing and Validating the Fuzzy Logic Expert System	83
Artificial Data	85
<i>Exemplary Range</i>	85
Upper and Lower Range of <i>Proficient</i>	86
Conclusion	87
CHAPTER FIVE-RESULTS.....	89
Research Question 1	90
Lack of a Perfect or Null Score Rating.....	90
Surface Views	91
Research Question 2	95
Lack of Variation	95
<i>Exemplary Range</i>	98
Upper Range of <i>Proficient</i>	100
Lower Range of <i>Proficient</i>	100
Conclusion	102
CHAPTER SIX-CONCLUSIONS AND IMPLICATIONS.....	104
Implications of a Fuzzy Logic Expert System.....	105

A Fuzzy Logic Expert System for Quantifying Teacher Performance	105
Distinguishing Between Performance Categories in the Fuzzy Logic Expert	
System.....	106
Distinguishing Between Performance Categories in the Traditional Method	107
Suggestions for Further Research	109
Reasons for Lack of Variability in New Teacher Evaluation Systems.....	109
Incorporating Fuzzy Neural Network Techniques.....	111
Confidential Principal Evaluations and Values on a Continuum.....	111
A Fuzzy Logic Expert System Using a Value-added model	112
Conclusion	113
REFERENCES	114
APPENDIX A: A Survey of TAPS Standards Affecting Teacher's Performance in	
Secondary Education	131
APPENDIX B: Definition and Description of Teacher Ratings.....	133

LIST OF FIGURES

Figure 1. Example of a matrix for classifying teacher effectiveness	13
Figure 2. Comparison of a well-defined and fuzzy set of tall people	16
Figure 3. Perceptions of two categories of speed	18
Figure 4. Figures of the triangle, trapezoidal, and sigmoid membership functions.....	20
Figure 5. Design of a fuzzy logic expert system.....	21
Figure 6. Fuzzification of the input variables	22
Figure 7. Application of the fuzzy operator in the antecedent.....	23
Figure 8. Apply implication method.....	24
Figure 9. Aggregate all outputs.....	25
Figure 10. Defuzzification	26
Figure 11. Surface view of performance evaluation	51
Figure 12. Example showing the relationship between the three tiers of the TAPS	65
Figure 13. Development of a classical expert system.....	71
Figure 14. Example of the membership functions of the input variable.....	76
Figure 15. Example of a decision matrix	77
Figure 16. The membership functions of the output variable	82
Figure 17. Portion of the design of the fuzzy logic expert system	83
Figure 18. Surface view of Teacher Performance versus Professional Knowledge and Assessment Strategies	92
Figure 19. Surface view of Teacher Performance versus Professional Knowledge and Assessment Uses	93
Figure 20. Surface view of Teacher Performance versus Professional Knowledge and	

Instructional Planning	94
Figure 21. Surface view of Teacher Performance versus Assessment Strategies and Assessment Uses	94
Figure 22. Surface view of Teacher Performance versus Instructional Planning and Instructional Strategies.....	95
Figure 23. Histograms of the distribution of <i>Exemplary</i> ratings of the traditional method	98
Figure 24. Histogram of the distribution of <i>Exemplary</i> ratings of the fuzzy method.....	99
Figure 25. Histogram of the distribution of outputs of the traditional approach	101
Figure 26. Histogram of the distribution of the outputs of the fuzzy logic method	102

LIST OF TABLES

Table 1. Example of Overall Summative Rating	66
Table 2. TAPS Final Ratings from Summative Scores.....	66
Table 3. Response Summary for Effect on Teachers' Performance	73
Table 4. Input and Output Variables Defined.....	75
Table 5. Subset of Rules Extracted from the Decision-making Matrix	78
Table 6. Parameters of the Membership Function of the Output Variable Teacher Performance	82
Table 7. Mapping between Crisp Output Ranges of the Fuzzy Logic Expert System and Linguistic Performance Values	83
Table 8. Frequency Table of Individual Standard Ratings of Participants (N = 510)	96
Table 9. Comparison of the Outputs of the Fuzzy Logic Expert System with the Summative Scores.....	97
Table 10. Comparison of Final Ratings for Traditional and Fuzzy Method.....	97
Table 11. Frequency of Exemplary Class Distribution of the Fuzzy Output.....	99
Table 12. Frequency of Class Distribution of the Fuzzy Output	102

CHAPTER ONE

INTRODUCTION

Researchers suggest that the most important factor in a student's education is first and foremost the teacher (Odden, Borman, & Fermanich, 2004; Goldhaber, 2009; Sanders & Rivers, 1996). If an effective teacher is the very core of student achievement, then the argument logically follows that an equally effective teacher evaluation system to reliably rate the level of teacher performance would in turn improve student learning. Yet, studies provide evidence that the current teacher evaluation systems are failing to provide a true picture of teacher performance by not accurately identifying the most talented teachers or those who need more support (Duffett, Farkas, Rotherham, & Silva, 2008; Kane, Rockoff, & Staiger, 2008; Toch & Rothman, 2008; Weisberg, Sexton, Mulhern, & Keeling, 2009).

Weisberg et al. (2009) in the New Teacher Project defined "the inability of our schools to assess instructional performance accurately or to act on this information in meaningful ways" as the Widget Effect (p.7). Results of their survey of 25,000 teachers provided evidence of leniency bias or over inflation of the highest rating category of teacher performance. Ninety-three percent of teachers received the highest rating, and only 3 of 1,000 teachers received "unsatisfactory" ratings (Weisberg et al., 2009). As with any profession, there is evidence of large variation in teacher performance (Kane et al., 2008; Sanders & Rivers, 1996). However, the inability to distinguish between different levels of teacher performance not only keeps schools from dismissing

ineffective teachers and recognizing exceptional ones, but it also prevents them from supporting growth among the vast majority of teachers who fall in the middle of the performance spectrum (Chukwubikem, 2012; Donaldson, 2012; Maslow & Kelley, 2012; Taylor & Tyler, 2012). The findings of Duffet et al.'s (2008) research suggested that teacher perceptions also supported these claims. In a survey of 1,010 K-12 public school teachers, Duffet et al. (2008) reported that only 26% of teachers found their most recent formal evaluation useful and effective.

Evaluation Reform in the United States

Since 2009, teacher evaluation has been a priority in efforts to reform public schools in the United States, partly in response to Race to the Top (RT3), a \$4.3 billion grant opportunity provided by the federal American Recovery and Reinvestment Act (U.S. Department of Education, 2009). Additionally, in the nonprofit sector, the Bill & Melinda Gates Foundation allocated \$290 million of grant money to support intensive partnerships for teacher effectiveness. An additional \$45 million was also allocated toward the 2009 Measures of Effective Teaching (MET) project to test new approaches to measure effectiveness in teachers (Bill & Melinda Gates Foundation, n.d.). The application guidelines for the Race to the Top (RT3) federal grant required states to develop more comprehensive teacher evaluation systems by using multiple measures of teacher performance, including student growth data, and by using multiple ratings in contrast to the traditional binary ratings of satisfactory or unsatisfactory to evaluate teachers (U.S. Department of Education, 2009). The results from multiple research reports of the MET project also advocate these requirements of multiple measures and

multiple ratings to identify and develop effective teaching (Cantrell & Kane, 2013; Joe, Tocci, Holtzman, & Williams, 2013; Kane & Staiger, 2012; Measures of Effective Teaching, 2013).

Teacher performance has become center stage in state education policy as well; to date, 19 states have received RT3 funding, and 34 states have modified their teacher evaluation systems (The Whitehouse, 2014). Common elements of states' teacher evaluation systems include a combination of teacher observations, primarily conducted by principals using multiple rating categories, and student achievement gains (Exstrom, 2013; Hull 2013; National Council of Teacher Quality, 2013; Shakman, 2012). Presently, forty-six states require the teacher evaluation system to use a model of multiple measures of teacher performance with one measure a type of student achievement, commonly referred to as value-added or a performance-based model (Exstrom, 2013; Hull, 2013).

Georgia Teacher Evaluation Reform

In 2010, Georgia applied for and was awarded \$400 million to implement its Race to the Top plan. As part of the state's teacher evaluation reform, Georgia implemented the Teacher Keys Effectiveness System (TKES). The goal of Georgia's Teacher Keys Effectiveness System (TKES) is "to provide teachers with meaningful feedback and support opportunities which lead to improved teacher performance and consequently, improved student outcomes" (Georgia Department of Education, 2013b, p. 11). Twenty-six school systems in Georgia began the pilot with a full implementation of the TKES system for the 2012-13 school year, and all the remaining school districts began the first

full implementation with the 2014-15 school year (Georgia Department of Education, 2014a). Certified teachers in Georgia receive a categorical Teacher Effectiveness Measure (TEM) based on documentation from two equally weighted components of the Teacher Keys Effectiveness System (TKES): the evaluator's ratings of a teacher's classroom performance based on the Teacher Assessment on Performance Standards (TAPS) and a student growth and achievement score (Georgia Department of Education, 2013a).

Classroom observations are conducted two times annually based on ten standards outlined in the Teacher Assessment on Performance Standards (TAPS). The TAPS are grouped into five domains, which follow: planning, instructional delivery, assessment of and for learning, learning environment, and professionalism and communication. Within each standard, evaluators rate teacher performance on a four-point rubric using the following terms: *Exemplary*, *Proficient*, *Needs Improvement*, or *Ineffective*, scored 3 to 0 respectively. The ratings for these standards are averaged to classify the overall rating of the teacher (Georgia Department of Education, 2013b).

While most states have adopted some type of measure based on student growth scores, classroom observations still remain an important part of the teacher evaluation process (Barrett, Crittenden-Fuller, & Guthrie, 2014; Hill, Charalambous, & Kraft, 2012; Hill & Grossman, 2013; Whitehurst, Chingos, & Lindquist, 2014). Whitehurst et al. (2014) suggests “nearly all the opportunities for improvement to teacher evaluation systems are in the area of classroom observations rather than in test score gains” (p. 2). A poll conducted by the National Council of Teacher Quality (2011) suggested that many states intend to rely exclusively on principals to conduct the classroom observations.

Problem Statement

Principal Evaluations

Earlier studies of teacher evaluation showed there is little correlation of principals' ratings of teachers with student achievement (Medley & Coker, 1987; Peterson, 1987, 2000). This is partially because reports suggested that principals award the highest rating to far too many teachers (Daley & Kim, 2010; Hill et al., 2012; Ho & Kane, 2013; Kimball & Milanowski, 2009; Weisberg et al., 2009). This would, therefore, restrict the evaluation score's range of variation and lower the overall rating-achievement correlations (Hill et al., 2012; Milanowski, 2004; Weisberg et al. 2009). Moreover, while principals have been found to accurately identify the teachers in their school with students who have the largest and smallest achievement gains, they were less able to distinguish among teachers between these extremes (Barrett et al., 2014; Batten, 2013; Jacob & Lefgren, 2005, 2008).

Combining Measures of Performance

While there is evidence that states should incorporate more measures to evaluate teacher performance, determining the best way to combine or aggregate multiple measures for an overall teacher performance measure is still a conflicted and disputed area of research (Cantrell & Kane, 2013; Hansen, Lemke, & Sorensen, 2013; Joe et al., 2013; Kane & Cantrell, 2010; Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2012; Measures of Effective Teaching Project, 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013; Partee, 2012; Yates, 2009). Hansen et al. (2013) stated "In doing so, the designers of these evaluation systems must make careful decisions not only about

selecting valid and reliable measures but also about how to combine performance ratings to minimize error and misclassification” (p. 1). Discussions of how to aggregate multiple pieces of information into a single measure of the overall level of performance are “easy to overlook or dismiss as ‘technical’ details” (Ammar, Bifulco, Duncombe, & Wright, 2000, p.264). While there is a body of research pertaining to how to aggregate value-added measures using teacher performance scores and student achievement scores (Ho & Kane, 2013; Kane, Taylor, Tyler, & Wooten, 2011), little research examines how to aggregate the individual ratings of classroom observations. If decisions about how to combine multiple measures of performance make little difference in whether or not teachers are identified as ineffective, then they should not warrant attention from stakeholders. On the other hand, if choosing the way these multiple measures are combined does indeed change the rating of a teacher and which teachers are rated as low-performers, the choice of aggregation method becomes an important consideration in a teacher evaluation system (Hansen et al., 2013).

Whenever individual outcomes are aggregated into a single measure of performance, information is lost. The choice in combining scores often leads to similar or even identical scores for the vast majority of teachers (Trstenjak & Donko, 2013). For example, if a teacher is rated as *ineffective* (score of 0) for one standard and *exemplary* (score of 3) for another, the average of 1.5 would round to same level of teacher performance as another teacher who received a score of *proficient* (score of 2) for both standards. Ammar et al. (2000) also highlighted the ability of a simple average to mask low areas of performance and the extent to which choices of how to aggregate individual results can affect the composite score when identifying low-performance schools.

Mihaly, McCaffrey, Staiger, & Lockwood (2013) suggest that by assigning weights to different performance measures, an evaluation system can establish and implement priorities into the evaluation system. This study examined how assigning weights established by evaluators could improve the way teachers were identified by different levels of teacher performance. Furthermore, this study assigned weights established by the evaluators of one particular school to reflect their priorities of teacher performance.

Rationale

A soft-computing approach known as fuzzy logic may provide a more effective and accurate approach to teacher evaluation (Bhosale & Kamath, 2013; Djam & Mishra, 2013; Atta-ur-Rahman, 2013). Fuzzy logic is a more intuitive approach aimed at managing imprecision in real-world applications and a powerful tool to efficiently represent the uncertainty and vagueness without complex, nonlinear mathematical models (Kumar, 2013; Massey, 2012; Mendel, 1995; Nykänen, 2006; Pappis & Siettos, 2005). The success of fuzzy logic in industry and other applications also provides evidence of its potential contribution to social science and education as well (Fourali, 1997; Singh, Gupta, Meitzler, Hou, Garg, Solo, & Zadeh, 2013). Advantages of the use of fuzzy logic provide a convincing rationale for this approach and are briefly discussed below. In order to further understand the advantages of fuzzy logic, a basic understanding of its theoretic foundation and a design of a fuzzy logic expert system are needed and are presented in Chapter 2.

Advantages of Fuzzy Logic in General

There are three advantages to a fuzzy logic approach in general. The first advantage is that a fuzzy logic approach is unique in its ability to simultaneously handle numerical data and qualitative linguistic knowledge (Mendel, 1995, Pappis & Siettos, 2005; Zadeh, 1975). Qualitative knowledge is often expressed using imprecise terms. Fuzzy logic incorporates reasoning that is approximate rather than exact with a set of rules. This tradeoff between significance and imprecision mimics how humans think and solve problems in order to make decisions (Kosko, 1993; Kosko & Isaka, 1993; Zadeh, 1975). Thus, fuzzy logic makes it possible to “distinguish between different shades of gray, similar to the process of human reasoning” (Massey, 2012, p. 6).

Secondly, the basis of fuzzy logic uses rules stated in natural language, which are computationally easier to solve than difficult nonlinear mathematical systems describing the same situation (Voskoglou, 2013). Traditional models for solving problems often fail to adequately record the complexity of the phenomena (Fourali, 1997). Complex systems tend to be high-dimensional, non-linear, and difficult mathematical systems to model or solve. The capability to incorporate dynamic system behavior “renders fuzzy logic systems almost indispensable for obtaining a more transparent and tactile qualitative insight for systems whose representation with exact mathematical models is poor and inadequate” (Pappis & Siettos 2005, p. 438).

Lastly, the wide range of applications already using a fuzzy logic approach of devising rules for vague expressions has proven to be useful for controlling various technology systems and modeling highly complex business problems, where success depends on fast and reliable decision-making (Durkin, 1990; Khan & Quaddus, 2004;

Singh et al., 2013; von Altrock, 1994; Voskoglou, 2013). The applications of expert systems, which were developed or adapted to fuzzy logic, are wide-ranging and growing in numbers. According to the Singh et al. (2013) report on the impact of fuzzy logic, there are presently 26 fuzzy logic research journals, 112,022 articles published on the theory or applications of fuzzy logic counted from two databases, and 16,898 patent applications in the United States related to fuzzy logic. In 1987, twenty years after Kosko's first paper on fuzzy logic was published, over two-thirds of the Fortune 1000 companies had already developed fuzzy logic expert systems to aid in the decision-making process (Durkin, 1990).

The Japanese were the first to apply fuzzy logic in the Nanboku line of the subway system in Sendai, Japan. Developed by Hitachi and put in operation in 1988, the fuzzy controller increased efficiency and stopping time and made the line one of the smoothest running subway systems in the world (Kosko & Isaka, 1993). Perhaps the first application of fuzzy logic control most tangible to the consumer has been in home and commercial appliances, specifically heating ventilation and air conditioning (HVAC) systems. By using fuzzy logic thermostats to control heating and cooling, the system saves energy by running more efficiently and keeping the temperature steadier than a traditional thermostat (von Altrock, 1994). A short list of other applications of fuzzy logic includes: aircraft control (Rockwell Corporation), cruise control (Nissan), automatic transmission (Nissan, Subaru), space shuttle docking (NASA), elevator scheduling (Hitachi, Fujitech, Mitsubishi), stock market analysis (Yamaichi Securities), TV picture adjustment (Sony), handwriting recognition (Sony Palm Top), video camera autofocus (Sanyo, Canon), and video image stabilizer (Panasonic) (Mendel, 1995).

There are also a growing number of fuzzy logic systems in non-engineering applications (Singh et al., 2013). The use and development of the expert systems has captured the attention of researchers in a number of fields and has been developed in a wide range of areas such as agriculture, chemistry, geology, medicine, space technology, etc. (Durkin, 1990; Fagan, 1978; Lemmon, 1986). Fuzzy logic even finds application in the entertainment world; the MASSIVE 3D animation system for generating crowds was used in making the Lord of the Rings (2001-2003) trilogy as well as the 2009 film Avatar (Ribeiro, n.d.). These examples of applications for fuzzy logic systems show that the principle of fuzzy logic can be used to model any continuous system, whether in the physical sciences or the social sciences or education (Kosko & Isaka, 1993; Wu, Tsai, Shih, & Fu, 2010).

Advantages to Fuzzy Logic in Education

There is also a host of literature that argues the value of applying fuzzy logic to the design of teacher evaluation systems (Khan, Amin, & Rehman, 2011; Cole & Pershitte, 2000; Fourali, 1997; Djam & Mishna, 2013; Atta-ur-Rahman, 2013). For educational research, fuzzy logic is a paradigm to introduce human subjectivity into a quantitative decision-making process and has the ability to model human knowledge and behaviors as they are, without ignoring their abstraction (Voskoglou, 2013). In the literature, three advantages of a fuzzy logic approach in education are listed and explained below: values on a continuum, alternate way to weight measures, and an alternate way to aggregate measures.

A fuzzy logic approach has great potential for improving teacher performance

evaluation by providing a clearer position of the continuum of teacher effectiveness when evaluation tools are not designed properly to do so (Prince, Koppich, Azar, Bhatt, & Witham, 2013). Conventional quantitative methods have clear-cut guidelines that lack the capability to be more sensitive to the relativity of the definition of each rating (Fourali, 1997). For example, in describing a person's height as tall, fuzzy logic can incorporate the distinction of somewhat tall versus very tall. Likewise, there is difficulty associated with classifying teachers who have just missed the defined criteria as contrasted to those who were further from the criteria (Kumar, 2013). Efforts have been made to improve teacher evaluation systems by extending traditional teacher effectiveness ratings beyond only two categories, pass or fail, to multiple categories (Kane et al., 2011). Fuzzy logic may prove helpful here. By describing teacher effectiveness as values on a continuum, or as a matter of degree, fuzzy logic may increase the accuracy of a teacher performance measure by taking into consideration the various uncertainties of human behavior and distinguishing between gray areas of the ratings. This would in turn allow evaluators to express measures in more flexible ways, which tend to be more accurate and reliable (Bhosale & Kamath, 2013; Fourali, 1997). Using rules incorporated in the fuzzy logic expert system to interpret the results of principal evaluations of teacher performance may improve a teacher evaluation system's ability to distinguish between performances of teachers in the middle of the continuum.

A second advantage of a fuzzy logic expert system in contrast to traditional evaluation methods is that it can assign different weights to individual performance measures when determining the composite teacher effectiveness measure (Trstenjak & Donko, 2013). As previously noted, results from simply calculating the mean of the

scores often lead to similar or identical performance measure scores for a vast majority of teachers (Trstenjak & Donko, 2013). Because the “correct weight” or input for teacher effectiveness in an evaluative system will never be exactly known, the input is weighted accordingly with the knowledge given by experts in this area of research. Fuzzy logic can also give the flexibility to reflect more personalized and local priorities by incorporating different weights for different measures (Partee, 2012).

Lastly, a fuzzy logic model can also incorporate and examine the way in which these multiple measures “come together” for a composite teacher evaluation score. Examining this effect “is as important as the properties of any one individual component” (Carlo, 2011, para. 5). Whenever an assessment decision involves multiple measures, conventional quantitative approaches lead to a cumulative loss of information when human judgment would account for these multiple aspects of performance and indicate differently (Ammar et al., 2000; Fourali, 1997; Yates, 2009). A set of rules for how various measures of a teacher’s performance evaluation are combined must go beyond basic operating formulas to arrive at the categories of effectiveness. The fuzzy logic model is a more advanced and sensitive alternative to the way scores from all categories are aggregated to produce a final score, which can result in a more realistic evaluation (Bhosale & Kamath, 2013; Trstenjak & Donko, 2013). Fuzzy logic is able to incorporate rules, such as those found on the tables below in Figure 1, or other decision-making strategies that better reflect human judgment than other existing quantitative approaches. For example, evaluators may agree that a score below a certain rating for any one particular criterion may automatically disqualify the teacher from being assessed as an effective performer. Such is the case in Figure 1 for a teacher with a Student Growth

Quality Standards Score	1	2	3	4	5
5	Partially effective	Partially effective	Effective	Highly effective	Highly effective
4	Ineffective	Partially effective	Effective	Effective	Highly effective
3	Ineffective	Partially effective	Effective	Effective	Effective
2	Ineffective	Partially effective	Effective	Effective	Effective
1	Ineffective	Ineffective	Partially effective	Partially effective	Partially effective
Student Growth Score					

Score of 1, with the only exception being a Quality Standards Score of 5.

Figure 1. Example of a matrix for classifying teacher effectiveness. Reprinted from National Council on Teacher Quality (NCTQ) (2011). Retrieved from <http://www.nctq.org>. Copyright 2011 by NCTQ. Reprinted with permission.

The rules that comprise a fuzzy logic expert system are applied in parallel rather than in sequence. A fuzzy logic procedure can be useful when multiple raters are involved to take into account all group members' ratings or classroom observations from multiple occasions (Fourali, 1997). Thus, fuzzy logic has the ability to model a system, which can involve multiple experts, and is even well suited when expert opinions represent conflicting opinions. Fuzzy logic also can take into account and keep track of any reservations or exceptionalities evaluators observe (Ingoley & Bakal, 2012).

Purpose of Study

While current teacher performance evaluation systems have been found to accurately identify the teachers with students who have the largest and smallest achievement gains in their schools, they were less able to distinguish among teachers in between these extremes (Batten, 2013; Jacob & Lefgren, 2008). Equally important, the

design of evaluation systems must incorporate new ways to combine performance ratings to minimize error and misclassification (Schochet & Chiang, 2010).

The purpose of this study was to examine to what degree a fuzzy logic expert system could elicit and quantify teacher performance using state teacher evaluation methods. Theoretical background is given as a foundation for fuzzy logic expert systems in order to understand their implications for teacher performance evaluations. Measures of teacher performance were examined in terms of the Georgia Teacher Keys Effectiveness System. Results from the proposed fuzzy logic expert system were analyzed to more accurately identify and distinguish levels of teacher performance. Finding the best ways to incorporate all the given information may ultimately lead to the truer teacher performance measure.

Research Questions

There are two main research questions that were addressed in this study:

1. How can a fuzzy logic expert system quantify teacher performance using the ratings from the principal evaluations?
2. How do the ratings of a fuzzy logic expert system identify and distinguish levels of teacher performance as compared to traditional state evaluation methods?

Conclusion

Because an effective teacher evaluation method is vital to student success, it must be based on valid measures that adequately and accurately capture the complexity of good teaching into the evaluation process. Most states have implemented reform in teacher evaluation strategies in order to diagnose and improve teacher performance but

have not incorporated a method capable of accurately representing qualitative factors and the relationships between them. In order to improve the decision-making process associated with teacher evaluation, there is a need for teacher evaluation systems to use a model that can handle the academic qualitative uncertainty of fine distinctions among teachers whose performance falls in a broad middle range. This study attempts to define such a model.

CHAPTER TWO

FUZZY LOGIC AND FUZZY SYSTEMS

Fuzzy Logic and Set Theory

In classical set theory, membership of an element to a set is binary: either it is included or it is not (Zadeh, 1965). In many situations it is difficult to apply such clear-cut boundaries to membership of one category from the next. The rules of categorization with hard boundaries fail to reflect the inherent imprecision of real life application, including human experience (Zadeh, 1983; Kosko, 1993). The example of a fuzzy set of tall people is commonly used to illustrate this concept (Kumar, 2013). When categorizing the height of people, it is illogical to label one person as short and another one as tall when the difference in height is the width of a hair. If the set of tall people is given as a hard boundary of a classical set, this is precisely what happens at the boundary, unable to account for the areas of different shades of gray (Figure 2).

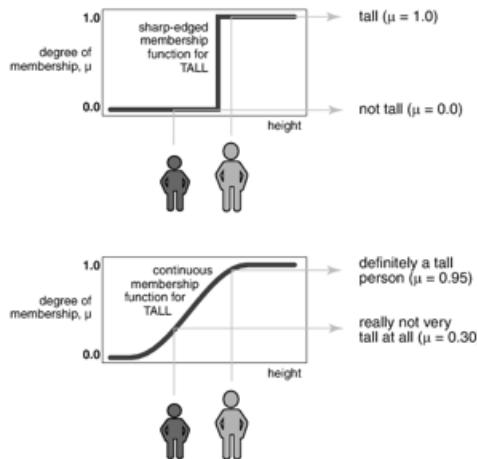


Figure 2. Comparison of a well-defined and fuzzy set of tall people. Reprinted from Mathworks, Inc., (2013). Retrieved from <http://mathworks.com>. Copyright 2013 by Mathworks, Inc. Reprinted with permission.

However, in 1965, Lotfi Zadeh, regarded as the father of fuzzy logic, proposed an alternative “fuzzy” logic and corresponding set theory. Fuzzy reasoning offers a way “to reply to a yes-no question with a not-quite-yes-or-no answer” (Mathworks, Inc., 2013, p. 2-4). Thus, the truth of any statement becomes a matter of degree, and an element’s degree of membership to the fuzzy set is given a value between 0 (definitely not a member) to 1 (definitely a member).

Zadeh (1965, 1975, 1983) defined linguistic, fuzzy variables as words quantifying the gradual transition of set membership such as from high to low or true to false. These quantifiers can capture the aspect of implicit human conversation. Such elements in a fuzzy set include words to refer to relative counts such as mostly or rarely. Zadeh (1975) illustrates this with the example, “Lynne is never late.” This statement would exclude Lynne from the set of late people entirely. Rephrasing the same statement as “Lynne is very rarely late” would allow it to be included in the fuzzy set of late people with a small degree of membership.

Membership Functions

The imprecise and vague linguistic inputs are transformed into precise numerical inputs by a membership function. This curve assigns a value to the degree an element belongs to the set.

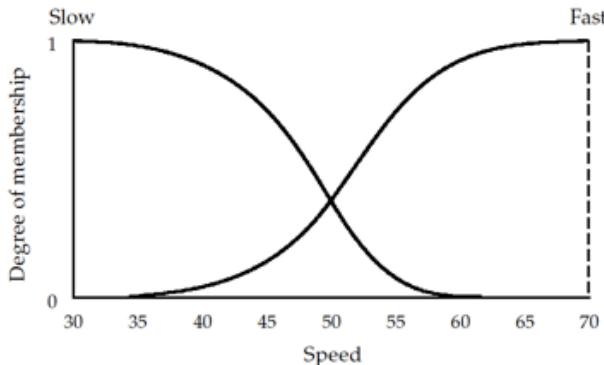


Figure 3. Perceptions of two categories of speed. Reprinted from
doi:10.1080/09500799708666923. Copyright 1997 by Fourali. Reprinted with
permission.

In Fourali's (1997) example, shown in Figure 3, a curve is given for the minimum speed of a car to be considered *fast*, with values ranging from 35 to 70 miles per hour. In the second curve, a curve is given for the maximum speed of a car to be considered *slow*, with values ranging between 30 miles an hour to 60 miles an hour. For example, suppose the speed of a car is 45 miles per hour. Forty-five miles per hour belongs to both sets, *fast* and *slow*, to a matter of degree. The degree of membership in the set labeled as *fast* is 0.2, and the degree of membership is 0.65 for the set labeled as *slow*. Although a membership function of a fuzzy set resembles a probability function, both operating over the same numeric range $[0, 1]$, the concept of a fuzzy set is nonstatistical in nature (Zadeh, 1965). Fuzzy logic does not estimate the likelihood of an event or an element having a certain property but the degree to which it has that property (Voskoglou, 2013).

The simplest membership function consists of straight lines to form a triangle membership function. This is frequently used in fuzzy logic applications to teacher performance evaluations (Chaudhari, Khot, & Deshmukh, 2012; Khan et al., 2011; Trstenjak & Donko, 2013). A second type of membership function is the trapezoidal membership function, which can be thought of as a truncated triangle function. The truncated part of the triangle is defined as a hedge. Hedges would linguistically represent the adverb modifier *rarely* in the previously discussed example, “Lynne is rarely late.” The truncated or plateau of the graph conceptually represents the small window of possibility that Lynne would arrive late. Other common membership functions are Gaussian membership functions with a smooth bell-shaped curve or the sigmoid function with asymmetric properties, important in certain applications (Figure 4) (Garibaldi & John, 2003; Zhao & Bose 2002).

Looking at Figure 4, each type of membership function is based on an input variable or rating on the interval 1 to 10, as shown on the x-axis. The y-axis represents the degree of membership the variable has to a set or category. In a triangle membership function, a rating of 6 is the only value that would reflect total membership ($y = 1$) to a set; whereas, in a trapezoidal membership function, all ratings between 5 and 7 would have degree of membership of 1 as well. Lastly, for a sigmoidal membership function, a nonlinear curve would represent the values from 1-6, and the ratings from 6-10 would have a total degree of membership. In applications to teacher evaluations, a review of studies shows examples of fuzzy logic systems using a variety of functions: triangle membership functions (Atta-ur-Rahman, 2013; Bai & Chen, 2008a; Trstenjak & Donko,

2013; Yadav & Singh, 2011), trapezoidal functions (Ahmad & Asri, 2013; Chaudhari et al., 2012) and Gaussian functions (Bhosale & Kamath, 2013).

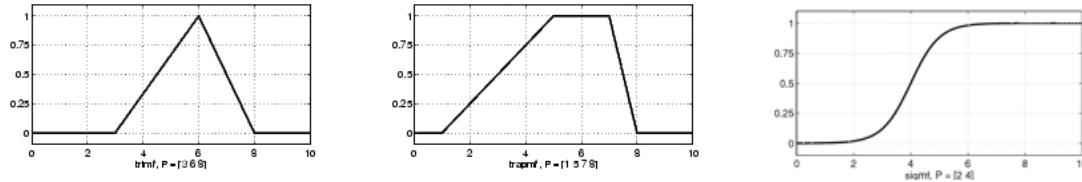


Figure 4. Figures of the triangle, trapezoidal, and sigmoid membership functions.

Reprinted from Mathworks, Inc., (2013). Retrieved from <http://mathworks.com>.

Copyright 2013 by Mathworks, Inc. Reprinted with permission.

Fuzzy Rules

The principle concept of fuzzy logic is the uniqueness in the way it maps input values of a fuzzy set to an output value in order to make a decision. The function in which it accomplishes this is a rule base of a list of if-then conditional statements or rules. In general, a rule has the form: If a is X, then b is Y. The if-part of the rule is commonly called the antecedent, while the then-part of the rule is called the consequent (Mathworks, Inc., 2013). The rule is interpreted as: b is a member of Y to the degree that a is a member of X. Consider the example, “If it is raining, then it is wet outside.” In other words, the last part of the rule, “it is wet outside,” is true only to the extent or degree the first part, “it is raining,” is true. Rules can also be weighted between 0 and 1 to give more importance to one rule over the others (Amin & Khan, 2009; Atta-ur-Rahman, 2013; Trstenjak & Donko, 2013).

Design of a Fuzzy Expert System

The first and critical step of the development of every fuzzy logic expert system is the construction of the knowledge acquisition process, in which human expertise of a subject domain is acquired to write the if-then rules of the fuzzy logic expert system. Once this is obtained, the design of a fuzzy logic expert system comprises four steps, as can be seen in Figure 5.

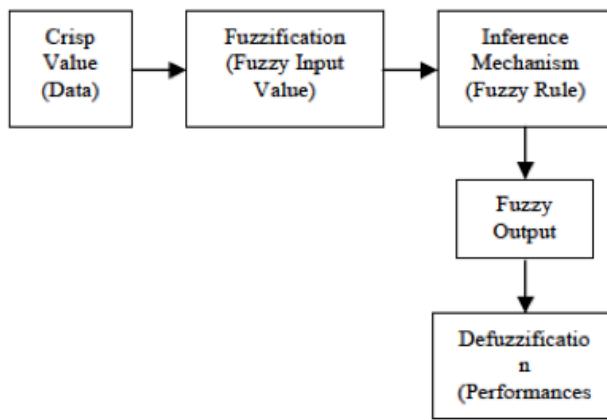


Figure 5. Design of a fuzzy logic expert system. Reprinted from Yadav & Singh (2011).

Copyright 2011 by Yadav & Singh. Reprinted with permission.

The first step is fuzzification. This means the list of input variables and their range of values is identified, and a fuzzy membership function is defined. The second step is to write rules using Zadeh's fuzzy set logic to combine the knowledge from the domain experts during the initial step to construct a list of fuzzy if-then statements. In the third step, the fuzzy logic expert system aggregates the input by making a set of inferences and associations between and among the input and output variables. The fourth and last step is to "defuzzify" the output variable with these inferences and associations. This means the resulting aggregated fuzzy region for the output variable is

transformed into crisp numerical values to reach a decision applicable to the problem at hand.

To illustrate the design of a fuzzy logic system, the same example will be used as in the Matlab Fuzzy Logic Toolbox tutorial (Mathworks, Inc., 2013), which illustrates the steps with the following question: “What is the right amount to tip a server in a restaurant as typically practiced in the United States?” (p. 1-12). While the standard rule of thumb to tip for a meal may be 15% of the total bill, the actual amount given can vary depending on the quality of food and service. Thus, for this example, there are two input variables defined, the quality of the food and the quality of service, and one output variable, the amount of the tip.

Step 1: Fuzzification of the input variables

First, the input values must be fuzzified. This basically means the degree to which the input variables belong to the fuzzy sets is determined by the membership function. The input variables, quality of food and quality of service, are each rated on the interval from 0 to 10. Figure 6 shows to what degree a food rated as an 8 in delicious qualifies in the set of the *delicious*. Given the definition of *delicious* by the shown membership function, the output corresponds to a value of 0.7.

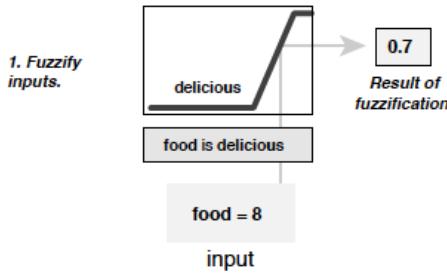


Figure 6. Fuzzification of the input variables. Retrieved from <http://mathworks.com>.

Copyright 2013 by Mathworks, Inc. Reprinted with permission.

Step 2: Assigning rules

Application of the fuzzy operator in the antecedent. In the tipping example (Mathworks, Inc., 2013), three rules are assigned:

1. If service is poor or the food is rancid, then tip is cheap.
2. If service is good, then tip is average.
3. If service is excellent or food is delicious, then tip is generous.

If the antecedent of a given rule has more than one part, the Mamdani's max-min fuzzy operator (*and* or *or*) is the most commonly applied (Mendel, 1995; Pappis & Siettos, 2005; Saleh & Kim, 2009). If the antecedent is joined by *and*, the minimum of the membership values is taken. On the contrary, if a rule is joined by *or*, the maximum is taken (Kosko, 1986). Figure 7 shows the antecedent of the third rule. The two input variables (service is *excellent* and food is *delicious*) from the fuzzy membership function result in values 0.0 and 0.7, respectively. Since the statement uses the fuzzy *or* operator, the maximum of the two values, 0.7, is used. In the implication process, the input for the implication process is a single number given by the antecedent, and the output is a fuzzy set that represents the consequent in each rule. In this case, the fuzzy set of a tip

considered generous is represented by a triangle membership function.

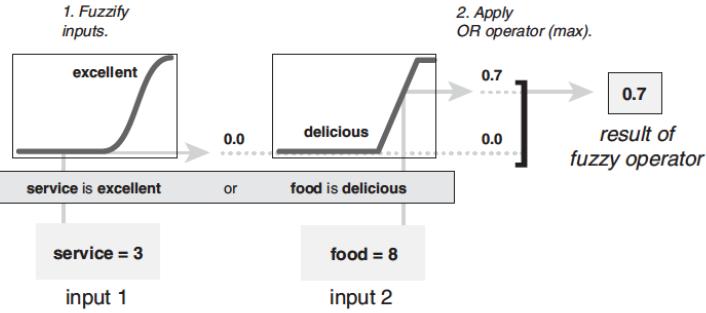


Figure 7. Application of the fuzzy operator in the antecedent. Reprinted from Mathworks, Inc., (2013). Retrieved from <http://mathworks.com>. Copyright 2013 by Mathworks, Inc. Reprinted with permission.

Figure 8 shows an example for the ratings of the input values as follows: service is 3, and food is 8. Then according to the membership functions of the input variables, the horizontal gray arrows in the antecedent depict the degree of membership that the service is considered excellent and the food delicious. Since the rule joins the variables Mamdami operator “or,” the maximum value is taken. Thus, the higher of the gray horizontal arrows truncates the triangle membership function, and the result of the implication method is the truncated output membership function shown in Figure 8 that describes the tip as generous.

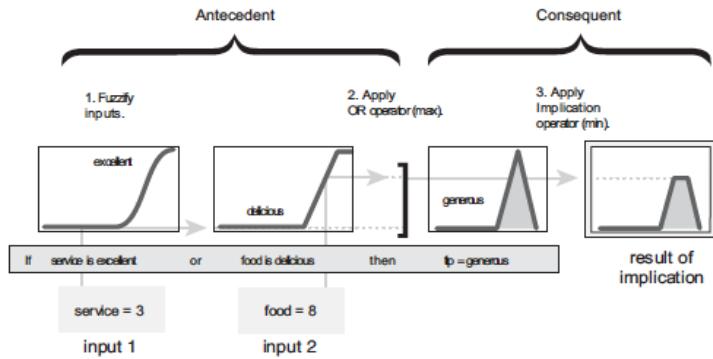


Figure 8. Apply implication method. Reprinted from Mathworks, Inc., (2013).

Retrieved from <http://mathworks.com>. Copyright 2013 by Mathworks, Inc. Reprinted with permission.

Step 3: Aggregate all outputs

As much as fuzziness helps the rule evaluation, the final desired output variable must be a single numerical value to reach a decision applicable to the problem at hand. The truncated output variables from each of the three rules in Figure 9 are shown in the right column. Each rule is aggregated and executed simultaneously by overlapping all results into a single fuzzy region as a result of the aggregation. This can be seen in the last picture of the right column, which keeps track of every output or tip value and its assigned weight.

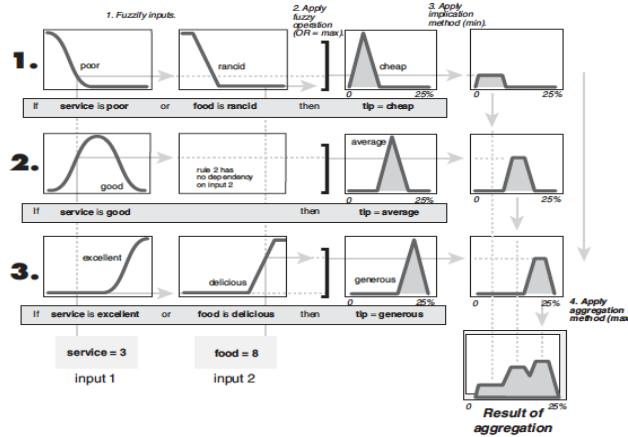


Figure 9. Aggregate all outputs. Reprinted from Mathworks, Inc., (2013). Retrieved from <http://mathworks.com>. Copyright 2013 by Mathworks, Inc. Reprinted with permission.

Step 4: Defuzzification

In the defuzzification process, the resulting aggregated fuzzy region is transformed into a crisp value as the output variable. This is most commonly done by the centroid or center of gravity method, which is simply the weighted average of the output membership function (Pappis & Siettos, 2005; Mendel, 1995). In Figure 10, the center of gravity is calculated as 16.7%, which indicates the suggested amount of tip for the tipping example problem (Mathworks, Inc., 2013).

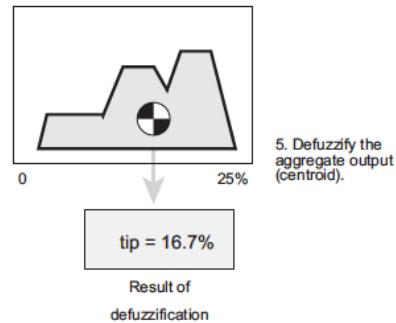


Figure 10. Defuzzification. Reprinted from Mathworks, Inc., (2013). Retrieved from <http://mathworks.com>. Copyright 2013 by Mathworks, Inc. Reprinted with permission.

CHAPTER THREE

REVIEW OF LITERATURE

Researchers have provided evidence of the validity of changes to teacher performance evaluation systems in the last several years by comparing traditional evaluation systems to value-added models (Kane et al., 2013). However, there is continuing confusion and debate over whether value-added measures are sufficiently reliable (Goldhaber & Loeb, 2013). Kane, Kerri, and Pianta (2014) provided evidence that teacher observations under certain conditions can provide an accurate picture of teacher performance. These measures are less accurate, however, when principal evaluations show an overly inflated number of teachers who earn the highest ratings (Weisberg et al., 2009). In addition, the manner in which teacher performance measures are combined and weighted in a teacher's total score is as important as the measure of any individual component (Carlo, 2011).

Teacher Performance Measures

Value-Added Models

The new direction of teacher evaluation using value-added models has become an active strand of research (Almy, 2011; Goe, Bell, & Little, 2008; Kane & Staiger, 2012; Measures of Effective Teaching, 2013; Milanowski, 2011; New Teacher Project, 2010). Logically, there should be a relationship between the effectiveness of a teacher and student learning, and teacher evaluation should center on measuring this relationship (Almy, 2011; Milanowski, 2011). Even the definition of teacher effectiveness is often

defined as “the ability to produce gains in student achievement scores” (Goe et al., 2008, p. 1). In Goe et al. (2008)’s research synthesis of approximately 120 studies that examined value-added measures of teacher effectiveness, the summary cautioned defining teacher effectiveness solely in this way for several reasons. Teachers are not exclusively responsible for student learning, learning is more than average achievement gains, and test scores are limited in the information they can provide.

Although logic could lead one to believe that better teachers lead to better student learning, questions still remain about the validity of value-added measures as an appropriate tool for identifying effective teachers (Carlo, 2011; Corcoran, 2010; Goldhaber & Loeb, 2013; Hallinger, Heck, & Murphy, 2014; Jacob & Lefgren, 2005; Jacob, Lefgren, & Sims, 2008; Kane et al., 2008; Schochet & Chiang, 2010). For instance, finding a strong relationship between student achievement gains and other measures such as classroom observations could be interpreted as parallel validation. If it were possible to know for certain that principal evaluation was the correct measurement of performance, a strong correlation between this measure and student achievement gains would validate a value-added model. On the other hand, if the same were possible in student achievement gains, a strong correlation would serve as validation of the principal evaluation. In theory, one affects the other, but does not perfectly predict the other (Daley & Kim, 2010; Kane & Staiger, 2012).

Several studies compared teacher effectiveness to value-added scores measured by observation of teaching practices, but the effects were modest and varied across the different sites (Gallagher, 2004; Milanowski, Kimball, & White, 2004; Milanowski, 2004). Milanowski (2004) examined the correlation in Cincinnati Public Schools

between teacher evaluation ratings and student achievement in third through eighth grades on state tests in reading, mathematics, and science. The teacher evaluation system was based on the standards derived from Danielson's (1996) *Framework for Teaching*, one of the most widely used instruments (Goe et al., 2008). In this framework, 16 performance standards were grouped into four domains: planning, learning environment, teaching strategies, and professionalism. For each standard, a rubric was used which described four levels of performance: *unsatisfactory*, *basic*, *proficient*, and *distinguished*. Two of the domains (learning environment and teaching strategies) were evaluated based on six classroom observations, and the other two domains were evaluated with a portfolio prepared by the teacher. An evaluator outside the school conducted four of the classroom observations, and only two were conducted by building administrators. Milanowski (2004) found small to moderate positive correlations for most grades, with a substantial amount of variation across subjects and across grades.

Gallagher (2004) found similar results in the relationship between 2nd through 5th grade student achievement scores in mathematics and language arts and teacher performance evaluation at Vaughn Elementary School, a charter school in the Los Angeles Unified School District. The Vaughn evaluation system was comprised of 12 domains, 10 of which were content specific. The evaluations based on a classroom observation of a trained peer in the particular content, an administrator, and a self-evaluation of the observed teacher were weighted evenly for an overall rating score. Findings suggested that the relationship was stronger in reading than mathematics with a substantial variation as well. These two studies were replicated while adding a third site, Washoe County School District in Nevada, and the results were comparable to those

obtained previously (Milanowski et. al, 2004). The average correlations within each site between the teacher evaluation scores and the residuals from the student achievement model controlled for student characteristics were 0.27. Although this value qualified as significant, Milanowski (2004) did suggest that correlations ranging from 0.3 to 0.4 imply only a small proportion (9% - 16%) of the variance in student achievement can be attributed to teacher performance. Some reasons that high correlations are unlikely to ever be found between teacher performance and student achievement are errors in teacher performance measures, errors in student performance measures, and the role of student characteristics such as motivation or skill in student learning (Milanowski, 2004).

There are reports, on the other hand, that found teacher effectiveness measures, were positively and significantly correlated with student achievement growth (Daley & Kim, 2010; Kane et al., 2013; Rockoff & Speroni, 2010). The Measures of Effective Teaching (MET) project, funded by the Bill & Melinda Gates Foundation, was a three-year project from 2010-2013 of approximately 3,000 MET volunteer teachers from seven districts in seven states (Measures of Effective Teaching, 2010). The research confirmed the use of multiple measures in teacher evaluation systems, of which one measure was student achievement gains (MET, 2010). Kane et al. (2013) used the data collected in the MET project to build a composite measure of teaching effectiveness consisting of three measures: student achievement, classroom observations, and student perception. The standards for classroom observations for the MET project were based on the University of Virginia's Classroom Assessment Scoring System (CLASS) framework, one of the two most widely used instruments (Goe et al., 2008; Pianta, La Paro, & Hamre, 2006). Students were randomly assigned to each teacher, and their achievement was tracked.

The student achievement measures were based on standardized tests in mathematics and language arts for fourth through eighth grade. Researchers compared predicted student outcomes to the actual differences that emerged at the end of the year. The findings suggested that measures of effectiveness did identify teachers who produced higher student achievement (Kane et al., 2013). Rockoff and Speroni (2010) found similar results in the extent to which teacher observations and student achievement scores of new teachers in New York City could be used to predict future student achievement gains. The Department of Education in New York City releases Teacher Data Reports each year to its teachers in fourth through eighth grades with a summary of teachers' value-added information. Their study suggested that evaluation systems have significant potential to help address the problem of low teacher quality (Rockoff & Speroni, 2010).

Other research suggests that the classification errors associated with value-added measures based on student test scores can be quite high (Corcoran, 2010; Goldhaber & Loeb, 2013; Schochet & Chiang, 2010). Schochet and Chiang (2010) addressed the extent to which misclassification of teacher performance in value-added models occurred. The report used results published from other research or reports to examine the likely error rates of measuring teacher performance with student test scores. When three years of data were used to calculate the value-added score, the type I and II errors, which lead to false negatives or false positives, were approximately 26 percent. The error for a false negative means that 1 in 4 teachers, who were *proficient*, were incorrectly identified as *needs development*. Similarly, the error for a false positive means that 1 in 4 teachers, who were *needs development*, were misclassified as *proficient*.

Corcoran's (2010) study, which also used data from the New York City Data Reports, described a percentile of a teacher's performance relative to other teachers with similar experience in the same grade and subject. Evidence suggested the margin of error in the teacher value-added estimate was 30 percentile points. This means the true rank of a teacher rated at the 60th percentile could be anywhere between the 30th and 90th percentiles. Although the classification error rates associated with value-added models in their studies were quite high, they may be lower than the error rate of classifications based on traditional measures of teacher effectiveness such as licensure status or years of experience (Goldhaber & Loeb, 2013).

Classroom Observations

One advantage classroom observations have over student achievement gains is the ability to take into account aspects of performance that extend beyond measured outcomes, such as how outcomes are achieved (Milanowski, Prince, & Koppich, 2007). However, research suggests that connections between value-added scores and observations scores are weak. In a report by Kane and Staiger (2012) about the MET project, ninety-nine raters scored one thousand videos of fourth through eighth grade mathematics lessons. The findings confirmed the connections between value-added model scores and observation scores as fairly weak. Thus, while committed to the use of student achievement scores to help measure teacher performance, the MET project acknowledged that teacher performance evaluation cannot solely be determined by student achievement gains. The majority of the project's published research appears to be focused on ways to improve the reliability of observation systems (MET, 2010;

Cantrell & Kane, 2013; Ho & Kane, 2013; Kane et al., 2013; Kane & Staiger, 2012; Mihaly et al., 2013; Wood, Tocci, Joe, Holzman, Cantrell, & Archer, 2014).

A wide body of literature gives recommendations for a more reliable observation system and confirms the use of multiple rating scales, multiple evaluations, and multiple observers to ensure evaluators' measures of teacher performance are reliable (Goldhaber & Loeb, 2013; Hill et al., 2012; Hill & Grossman, 2013; Kane & Cantrell, 2010; Milanowski, 2007, 2011; New Teacher Project, 2010; Tyler, 2011; Wang, Wong, & Wong, 2010; Weisberg et al., 2009; Wood et al., 2014). Traditionally, evaluation ratings were binary, satisfactory or unsatisfactory, which could not give a clear picture of where teachers fell in the spectrum of performance (Kane & Staiger, 2012; New Teacher Project, 2010; Stronge, Xu, Leeper, & Tonneson, 2013). The scale of the instruments most states now use have four or five ratings. Georgia has four ratings for teacher performance: *Exemplary, Proficient, Needs Development, and Ineffective* (Georgia Department of Education, 2013b). Hill et al. (2012) reported that observations were most effective when conducted multiple times per year due to widely documented variance in observational scores from lesson to lesson, and one observation per year is unlikely to provide a fair or representative sample of teacher performance.

The results of the Wang et al. (2010) study confirmed the benefits of using multiple evaluators and indicated that raters with different goals gave different ratings. An ANOVA was used to examine the effects of a rater's goals on rating scores with the four most common rater goals: different levels of performance identification, harmony, fairness, and motivation (Wang et al., 2010). Raters deflated ratings for high performers to achieve the fairness goal, and they inflated their rating more for low performers to

motivate them (Wang et al., 2010). Incorporating multiple evaluators can also allow classroom observations to be subject-specific and include content experts in the observation evaluation (Hill & Grossman, 2013; Tyler, 2011). Hill & Grossman (2013) recommended multiple evaluators, particularly in a high school, because principals do not always have adequate knowledge of the subject being taught.

Principal Evaluations

Despite evidence that a reliable observation system should include multiple evaluators, recent trends in teacher evaluation seem to suggest the role of principals, as evaluators, will continue to expand in the coming years (Hill et al., 2012; Kimball & Milanowski, 2009). The quantity of research on classroom observations and principal evaluation pales in comparison to the research focused on the validity of value-added measures (Whitehurst et al., 2014). The reports, in respect to a principal's accuracy in predicting teacher performance, are somewhat mixed with correlation coefficients between teacher evaluations and value-added scores being weak to moderate (0.17 – 0.49), depending on data and methodology (Batten, 2013; Prince et al., 2013).

In an earlier study by Medley and Coker (1987), the accuracy of principal evaluations was measured for 46 elementary school principals. The researchers examined the relationship between principal judgments and student achievement gains, which were estimated from students' pretest and posttest scores of the same year in standardized achievement tests in reading and mathematics. The researchers reported that the correlation between the principals' judgment and student achievement test scores was practically zero. Similarly, in a review of the literature, Peterson (2000) found that

principals were reported to not be accurately evaluating teacher performance, and teachers had little confidence in evaluation results. Reasons for the inaccurate evaluations ranged from content expertise or lack of content knowledge to bias to desire to preserve school harmony or not upset working relationships (Peterson, 2004).

Later research suggested that principal evaluations could be an effective measure for teacher performance. Harris and Sass (2009) found a positive correlation between principal evaluation ratings and value-added student achievement scores in a study of a Florida school district. The researchers conducted a factor analysis of 11 individual teacher characteristics and an overall assessment of the teacher, on a scale of 1 to 9, rated by principals during an interview with the researchers. These ratings were compared to the Stanford Achievement Test achievement gains for students in second through tenth grades over a five-year period. Harris and Sass (2009) found that adding principals' subjective ratings to prior value-added scores improved the estimated future value-added measure. In addition, Rockoff, Staiger, Kane, and Taylor (2011) examined the relationship between 223 principal evaluations in New York City elementary and middle schools and student achievement scores on fourth through eighth grade mathematics and English standardized tests. The study claimed a strong effect size of 0.23 using a series of linear regressions in a Bayesian learning model.

Despite some evidence of principals' ability to identify effective teachers, principal evaluations are frequently lenient with the vast majority of tenured teachers at the highest possible point on the scale (Daley & Kim, 2010; Jacob & Lefgren, 2008; Weisberg et al., 2009). A teacher performance evaluation system that is overly lenient and inflated cannot differentiate between the effectiveness of teachers in a meaningful

way to stakeholders and teachers themselves. Weisberg et al.'s (2009) report in the New Teacher Project reflected survey responses from 1,300 administrators and 15,000 teachers in 12 districts and 4 states. The results from observations indicated that administrators rated less than 1% of teachers as unsatisfactory. Despite uniformly satisfactory ratings, 81% of the administrators indicated there was a poor performing tenured teacher in their school, but over half of the districts during the time of the study did not dismiss a single teacher for poor performance.

In a study by Jacob and Lefgren (2008), it was found that "principals are quite good at identifying teachers whose students make the largest and the smallest standardized achievement gains in their schools but less able to distinguish between teachers in the middle of the distribution" (p. 29). Elementary school principals were asked to rate a sample size of 201 teachers on a scale from 1 to 10 of overall teacher effectiveness and also assess specific teacher characteristics such as work ethic, parent satisfaction, and positive relationship with colleagues. Using conditional probabilities through Monte Carlo simulations, the researchers compared the probability a teacher rated by the principal in the top category would also have a value-added measure in the top category with the expected probability if the principal ratings were randomly assigned. The researchers reported a difference of 38 percentage points in the two probabilities for teachers rated in the top category and a similar difference for the bottom category. However, the difference of 16 percentage points was considerably smaller in the middle category. This evidence supports their claim that principals are less able to distinguish between teachers in the middle of the distribution of teacher performance. It is important to note that the principal evaluations were completely confidential and not used for any

type of teacher performance measure.

Batten (2013) found similar results, reporting that principals had difficulty distinguishing between teachers whose performance fell in the middle of the broad spectrum. Using evaluation data on 26,260 North Carolina teachers, the researcher reported the evaluation scores were high and narrowly distributed. On a scale from one to five, more than half of the teachers received a composite score of 4 or higher. In addition, the average evaluation score of the 100 teachers with the lowest student growth scores was 3.4 while the average evaluation score of the 100 teachers with the highest student growth score was only a 4.1 (Batten, 2013).

Ho and Kane (2013) conducted an analysis of 11,800 North Carolina math teachers who received a formal principal evaluation using Danielson's (1996) *Framework for Teaching*. They reported that principals' ratings were compressed, and principals almost exclusively used the middle two of the four available ratings when identifying the performance of teachers. For any item evaluated, an average of 5% of teachers were in the bottom category, and just 2% were in the top category. With the vast majority of scores in the middle, only 0.1 point difference in a score was needed to move a teacher 10 points in percentile rank (Ho & Kane, 2013).

The report by Barrett et al. (2014) also supported a claim of limited variance and compression of ratings in principal evaluations on each of the five standards of teacher performance for all school districts in North Carolina. Principals demonstrated the tendency to assign only two ratings on the five-point scale for each standard. Although any one standard did not provide a large variance in performance, the composite measure, or mean, of the five ratings for each teacher provided more variance. Additionally, the

method in which Barrett et al. (2014) addressed this difficulty subtracted the average rating value for all teachers in the school from each teacher's composite measure to create a school-mean centered composite rating, which created a more normal distribution of scores (Barrett et al., 2014).

However, evidence by Barnett, Rinthapol, & Hudgens (2014) based on approximately 46,000 observations, suggested the ratings of the System for Teacher and Student Advancement (TAP) teacher performance evaluation framework developed by the National Institute of Excellence in Teaching was not all skewed at the top of the scale as in a typical evaluation system. The results provided evidence of a more normal distribution of teacher performance that more closely matched the assumed true distribution of teacher effectiveness (Barnett et al., 2014).

Additionally, there has been evidence of ratings with more variability in the District of Columbia Public Schools. In 2010, a teacher evaluation system, IMPACT, was introduced, which linked teacher performance to merit pay as well as the possibility of dismissal. Teachers rated as “highly effective” received substantial compensation while hundreds of teachers rated as “ineffective” were dismissed. This system also showed more realistically distributed data. In 2011-12, 16 percent of teachers were rated as “highly effective,” and 15 percent of teachers are rated as either “ineffective or “minimally effective” (Dee & Wyckoff, 2015).

In Georgia, there are different implications for teachers classified in the middle two categories, *Proficient* and *Needs Development*, even when there is little statistical difference between them. Teachers classified as *Needs Development* are required to have a Professional Development Plan (PDP) and be monitored by the building administrator,

while teachers whose ratings are only a little statistically different, are classified as *Proficient* and are not required to have a PDP (Georgia Department of Education, 2013b). For this reason, it is crucial that a teacher performance evaluation system accurately distinguish between these classifications, especially when multiple measures are used in the calculation of a teacher effectiveness measure.

Scoring Methods

While research has drawn attention to the use of multiple measures in teacher performance evaluation systems, there is limited empirical research to guide policy makers on the best way to combine or aggregate these measures to achieve district goals or on the properties in which measures come together to give the overall teacher effectiveness score (Cantrell & Kane, 2013; Kane & Staiger, 2012; Measures of Effective Teaching Project, 2013; Mihaly et al., 2013; Partee, 2012; Yates, 2009). Although these details might be seen as merely statistical or technical details, the decisions of creating a robust scoring design have important consequences for the accuracy and reliability of teachers' performance ratings (Hill et al., 2012).

Hansen et al. (2013) compared three commonly used approaches combining multiple measures of overall teacher performance: the numeric approach, the hybrid approach, and the profile approach. Although all models naturally introduce error or bias that did not exist in the individual performance measures, the researchers used an error-minimizing approach to explore the evaluation system's ability to reliably identify high and low-performing teachers and minimize misclassification error across different effectiveness categories. In each approach, the overall teacher effectiveness measure was

compared to a target criterion. The researchers defined the target criterion as the distribution of teachers into their “true” performance categories based on a long-term value-added model in the following way: *Ineffective* teachers, 10th percentile and below, *Needs Development*, 11th to 20th percentile, *Proficient* teachers, 21st to 80th percentile, and *Exemplary* teachers, 20th percentile and above. In their findings, the *numeric* approach, which calculates the weighted average of the raw scores of each indicator, introduced the least amount of bias, which was not statistically different from the target criterion (Hansen et al., 2013).

The error on the overall combined measure increases in the hybrid approach, which rounds the score of each measure to a whole number corresponding to a categorical rating. These whole numbers are then averaged for the overall teacher effectiveness measure. This rounding introduced a small bias that favored teachers (Hansen et al., 2013). For example, when the values of the category ratings *Needs Development* (value of 1) and *Proficient* (value of 2) are averaged, the resulting value of 1.5 would round to a categorical rating of *Proficient*, in favor of the teacher. Lastly, the profile approach, like the hybrid approach, categorizes teacher performance, but the profile approach rounds to a category of performance in multiple steps. Findings confirmed the potential for bias in this method further increased in favor of the teacher (Hansen et al., 2013).

Georgia uses a profile approach in combining the TAPS and student achievement measure. The summation of the individual TAPS measures are averaged and then assigned to a category of teacher performance. The same categorization is done for student achievement. Lastly, the integer values for the respective category are combined

to determine the teacher effectiveness measure. Thus, as defined in the profile approach, rounding the scores in multiple steps introduces the most amount of bias.

Weighting criteria

Many states and districts are combining multiple measures into a single index to provide feedback to teachers and to support decision-making (Hull, 2013). There is no formula on how much weight should be placed on each measure of effective teaching. Researchers suggest that the weight given any measure be determined by local priorities, and that states and districts test different weights to determine how changes affect the total performance in order to best represent particular contexts (Partee, 2012).

Cantrell and Kane (2013) built and compared four different weighted models to study the implications of different weighting schemes to produce different outcomes. The overall teacher performance score was a composite of student achievement gains, classroom observations, and student surveys from the data of the MET project. While the goal of the research was not to suggest a specific weighting scheme but to explore the difference of outcomes when choosing weights, the report recommended a balanced approach of assigning 33 to 50 percent to each weight (Cantrell & Kane, 2013).

In the report of Mihaly et al. (2013), a target criterion similar to the study by Kane et al. (2013) was assumed to estimate the optimal weights for each measure using data from the MET project. The weights for each measure were determined by aligning the overall score to closely resemble the target criterion. Researchers recommended weighting schemes of more equal weights as better predictors of teacher performance (Mihaly et al., 2013).

While there is research examining the way multiple measures are weighted in a value-added model, there has been only one study to date that has examined how individual criteria within an observation affect the overall score (Kane et al., 2011). In Kane et al.'s (2011) study, they used a statistical technique of principal component analysis, which "identified the smaller number of underlying constructs that the eight different standards of practice are trying to capture" (Kane et al., 2011, p. 57). The results showed that three indices explained 87% of the total variance of the eight standards. The first index was the teacher's average score across all eight standards. The second index examined two of the standards: classroom management vs. instructional practices. A teacher who was more skilled at classroom management than instructional activities received a higher score than a teacher who was more skilled in instructional activities than classroom management. The report claimed to discern relationships between more specific teaching practices in their effort to find a relationship between classroom observations of teachers and student growth data.

In the last year, some states have identified classroom observation standards or domains, which they weight more heavily than others (Connecticut Department of Education, 2014; Daley & Kim, 2010; New Jersey Department of Education, 2014). The weighting scheme among the states varies, however. For example, the weighting scheme in New Jersey has a balanced approach: Planning 20%, Environment 30%, Instruction 30%, and Professionalism 20%. In contrast, the weighting scheme in Connecticut heavily focuses on one domain: Planning 10%, Instruction 75%, and Leadership 15%. The weighting scheme in Tennessee, as with TAPS, accounts for differentiation of teacher roles between career, mentor, and master teachers (Daley & Kim, 2010). As part

of their roles, mentor and master teachers assume greater responsibilities outside of the classroom. The weighting for mentor and master teachers accounts for leadership activities through reduced weighting in instruction and increased weighting on leadership responsibilities (Daley & Kim, 2010). Georgia gives equal weight to each of the ten TAPS standards.

With these changes in order to advance teacher evaluation reform, there is a need for research to ensure accuracy of the teacher evaluation systems implemented as the complexity of their design increases. In addition, researchers are faced with the challenging task of creating teacher evaluation systems that can effectively distribute teachers across a continuum of effectiveness and create a stronger focus on how multiple measures affecting teacher performance should be aggregated to portray the overall score.

The success in other industries of a soft-computing approach known as fuzzy logic provide evidence of its potential in providing a more effective and accurate approach to teacher evaluation (Bhosale & Kamath, 2013; Djam & Mishra, 2013; Fourali, 1997; Atta-ur-Rahman, 2013; Singh et al., 2013). In order to further understand the advantages of fuzzy logic, a basic understanding of its theoretic foundation and the design of a fuzzy logic expert system are needed and are presented in the following section.

Fuzzy Logic Expert Systems in Education

Research regarding the use of fuzzy logic to education is almost exclusively applied in two areas: teacher evaluation and student assessment. Although this study will apply fuzzy logic to teacher evaluation, articles examining its use in student assessment

still have significant contributions to the application of fuzzy logic to education and the design of the fuzzy logic expert system, which can be applied to this study. For that reason, significant findings from studies of fuzzy logic applications in student assessment are included in the review of literature.

Law (1996) gives three reasons why the use of fuzzy logic in student assessment is appropriate. To begin, the observed scores are indeed imprecise or consist of vague data and can better be represented by fuzzy sets. Secondly, student scores fluctuate slightly from each assessment, and multiple scores are needed to provide an accurate picture of student performance. Lastly, the definitions of different ratings of performance are also vague data. These reasons can also provide an argument for why the use of fuzzy logic in teacher performance evaluation is appropriate. There are several aspects of the design of the studies in the review below that can be related to teacher performance and the design of this study: multiple measures, qualitative-type measures, evaluator bias, and difficulty of content or other external factors.

Review of Studies Using Fuzzy Logic Applied to Student Assessment

Student-centered learning has the need to incorporate multiple measures for student assessment just as teacher performance evaluation systems need to incorporate multiple measures into the overall teacher effectiveness score. Early work focusing on the application of fuzzy logic student grading systems in education showed potential for a fuzzy logic expert system to aggregate multiple test scores in order to produce a single score or grade (Bai & Chen, 2008a, 2008b; Biswas, 1995; Chang & Sun, 1993; Chen & Lee, 1999; Chen & Wang, 2009; Chiang & Lin, 1994; Echauz & Vachtsevanos, 1995;

Ma & Zhou, 2000; Nykänen, 2006; Semerci, 2004; Weon & Kim, 2001). Since then, this research has expanded beyond simply aggregating multiple scores to a large menu of subject measures of student learning such as teacher leniency, qualitative-type answers, difficulty of questions, and motivation (Ahmad & Asri, 2013; Bjelica & Rankovic, 2010; Huapaya, 2012; Ingole & Bakal, 2012; Jamsandekar & Mudholkar, 2013; Kannemeyer, 2005; Kao, Lin, & Chu, 2012; Mossin, Pantoni, & Brandão, 2010; Musavian & Ahmadi, 2013; Saleh & Kim, 2009; Singh & Pratap, 2011; Sripan & Suksawat, 2010; Srivastava, Rastogi, Srivastava, Saxena, & Arora, 2010; Taylan & Karagözoglu, 2009; Voskoglou, 2013; Yadav & Singh, 2011; Yadav, Ahmed, Soni, & Pal, 2014).

Multiple Measures. Yadav and Singh (2011) claimed a fuzzy logic approach to student assessment was able to make more sensitive types of distinctions between students by comparing 1st and 2nd semester examinations of third-year students in the Department of Computer Science in Varanasi, India. For example, consider two scores of two different students for 1st and 2nd semester: 90 and 70 for student 1 and 70 and 90 for student 2. The average would indicate the same level of student success while the reality of the situation indicates one student is improving while the other is not. Yadav and Singh (2011) claimed evaluation with fuzzy logic has greater flexibility and reliability when compared to the classical method. The fuzzy logic evaluation system revealed differences in the performance values. Success in this study was defined as a performance value above 50. For scores below 50, the performance value of the fuzzy logic system was smaller than the classical one; however, for scores above 50, the performance value was larger than the classical method. For example, a student with a

first semester score of 34 and a second semester score of 60 is unsuccessful in the classical method (47) but successful in the fuzzy logic system (63).

Bjelica and Rankovic (2010) introduced new software using fuzzy logic to evaluate and assess an unidentified sample of students in Serbia in their mathematics knowledge. The model incorporated all activities of students, including attendance of lectures, interaction during the class hours, homework, school projects, and scores on the midterm and final exams. When calculating the result from multiple measures, each activity was assigned a weighted percentage and defuzzified using the centroid method. In comparison to traditional methods of student assessment, this method did not measure the negative results, and it allowed students to go back and improve that value. From the software, the teacher was not only able to make recommendations that could improve the student's grade but also able to predict what the grade could be at the end of the term.

Mossin et al. (2010) proposed a fuzzy logic system to evaluate students in a distance learning industrial automation course in Sao Paulo, Brazil. Mossin et al. (2010) acknowledged an inadequacy of the traditional evaluation method to overlook a student who may fail the course but perform extremely well in one particular area. The study considered three scenarios where the system divided the students' evaluation into three areas and used specific weights for each to give the overall score of the student. The researchers claimed to have designed an evaluation system to help teachers in the industrial automation area identify students' strong and weak points by allowing the qualitative aspects to have more priority than the quantitative ones.

Motivation. Fuzzy logic systems have also been used to analyze and compare the levels of students' motivation and anxiety. Srivastava et al. (2010) proposed a fuzzy

logic system for analyzing and comparing students' motivation and anxiety at the Krishna Institute of Engineering and Technology in Ghaziabad, India. Twenty-four subjects in their third year at the university from two different disciplines (Computer Science and Engineering and Electronics Engineering) were given the Sinha's Comprehensive Anxiety Test before and after an exam. The researchers used a fuzzy logic expert system with anxiety as the input variable and motivation as the output variable to compare motivation levels between gender and the two academic disciplines. By employing a paired t-test, the researchers reported there was no difference in the motivation levels between gender, and motivation was slighter higher for the students in Electronics Engineering. The researcher claimed by using fuzzy data instead of raw data, they reduced uncertainty in imprecise information. However, there was no comparison made between using fuzzy and raw data.

Qualitative-type answers. Student-centered learning has the need to incorporate more qualitative or subjective-type measures of student assessment as does teacher performance evaluation. In an earlier study, Kannemeyer (2005) developed an instrument to assign a value for subjective types of questions to access students' understanding in mathematics from their answers to open-ended questions in a university calculus course. Using an empirical example of three students' semester exam, the findings suggested that incorporating fuzzy logic in the framework provided an adequate instrument for measuring student understanding. In Voskoglou' (2013) model, fuzzy sets represented three qualitative student characteristics: knowledge of the subject matter, problem solving skills, and analogical reasoning abilities in order to characterize student performance at the Technological Educational Institute of Patras, Greece. Using fuzzy

logic with a technique for assessing the deviation of a student's knowledge with respect to the teacher's knowledge, the researchers reported that fuzzy logic offered a richer assessment of the students' performance than a crisp numerical score or grade assigned to the student's success.

Difficulty of Questions. Several studies have considered the complexity, importance, and difficulty of questions in a student assessment (Bai & Chen, 2008a; Saleh & Kim, 2009; Weon & Kim, 2001). In these studies, the complexity of the question was weighted by the amount of time spent to answer it, and if a question was rated as having relatively high [low] importance or difficulty, the weight of the response accuracy increased [decreased]. In Bai and Chen's (2008a) method, they considered the difficulty, importance, and complexity of questions in the students' assessment. Using an empirical example of ten students each answering five questions, the researchers claimed the method provided a useful way to distinguish the ranking order of students with the same score, consistent with the findings of the earlier studies cited.

Teacher Lenience. An effective student assessment should aim to resolve the subjectivity of the teacher who assigns the grade, as should a teacher evaluation system aim to resolve the subjectivity of the evaluators. In one of the earliest works of fuzzy logic applied to student assessment, Biswas (1995) developed a fuzzy logic system constructed with membership functions to transform the grade of a student obtained from a strict-type or a lenient-type teacher into a normal-type grade. Cheng and Lee (1999) extended Biswas' method and claimed their method executed faster without performing mathematically complicated matching functions. This method was later again extended by Bai and Chen (2008b). The researchers presented a method for automatically

constructing the membership function of a lenient-type grade, strict-type-grade, and normal-type grade when evaluating teacher performance. For example, if a strict-type teacher assigned a grade of 36 to a student, the normal-type grade membership function would reflect this grade as 49. Using an empirical example of ten teachers evaluating the same five answers, the researchers claimed that by using fuzzy reasoning to infer the scores of the students, the system evaluated the scores in a fairer manner.

Ingoley and Bakal (2012) used a fuzzy logic expert system to combine three different qualitative aspects: student's degree of confidence (measured by the time taken to attempt the questions), question's complexity and importance, and evaluator's leniency, to write one composite student score. The researchers used the methods of Wang and Chen (2008) for degree of confidence, the method of Saleh and Kim (2009) for a question's complexity and importance, and the method of Bai and Chen (2008b) for evaluator's leniency, constructing membership functions for lenient, strict, and normal-type grades. The researchers provided an example of the grades of three students and compared the grades of the traditional approach to the fuzzy logic approach. The researchers claimed the examples supported that the fuzzy logic approach was effective in evaluating subjective types of answers fairly and transparently because it accounted for these differences.

The study of Kao et al. (2012) is the only known research to date that statistically compared a fuzzy logic expert system to a traditional method in student assessment or teacher performance. Kao et al. (2012) used a fuzzy logic expert system to diagnose student achievement and provide personalized and adaptive learning for learners to strengthen their understanding of a concept. Fifty-two students enrolled in a course at a

university in Taiwan were equally grouped into a control and experimental group. In the control group, the students completed a remedial learning activity, which provided personalized learning guidance based on the accuracy of questions answered. In the experimental group, the students completed a remedial learning activity that used a fuzzy logic expert system to provide personalized suggestions and adaptive learning materials. The fuzzy logic system not only took the accuracy rate into account as a diagnostic factor, but the difficulty of the question, the confidence level, and length of time to answer the question. After the pre-test, all students completed the personalized learning guidance recommended by each system before taking the post-test. The study showed in a paired t-test that there was no significant difference ($p > 0.05$) between the control and experimental group on the pre-test but significant difference ($p < 0.05$) between the groups on the post-test, indicating that the fuzzy logic expert system was more effective in evaluating and improving students' learning achievement.

To date, there has only been one study by Jamsandekar and Mudholkar (2013) that compared the distributions of outputs of a fuzzy logic expert system with a traditional one. In the study by Jamsandekar and Mudholkar (2013), the researchers determined student performance of 51 first-year students in the Department of Computer Science at Shivaji University in Kolhapur, India using a fuzzy logic expert system and compared it with a traditional evaluation method. The traditional method used the average percentage of four theoretical subject papers and one lab/practical paper to classify student performance. In the fuzzy logic expert system, these five input variables were assigned three membership functions named low, medium, and high, and the output variable was defined with five membership functions named poor, below average,

average, above average, and good. The surface view of the performance evaluation space can be seen in Figure 11. The two input variables, lab percentage and combined average percentage of the four theoretical subject papers, are depicted on the x and y-axis, and the student performance output value is plotted on the z-axis. It can be seen from Figure 11 that the output value of performance for input values in the higher and lower ranges, labeled *Good* and *Poor*, remained stable. The output value was more affected for the input value range between 40-75, labeled *Conflict Area of Evaluation*. For this range of values, the surface view showed distinct areas of performance with smooth transitions from one category to the next for this region. The researchers claimed that this smooth distribution helped to identify students lying at the border of two class distributions, where these spaces have previously imposed challenges to the evaluator.

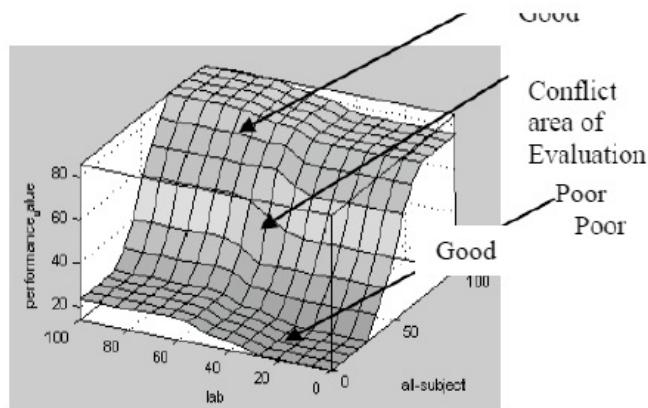


Figure 11. Surface view of performance evaluation. Reprinted from www.ijcse.org/attachments/File/v3i2/B1477053213.pdf. Copyright 2013 by Jamsandekar & Mudholkar. Reprinted with permission.

In the traditional method, the distribution of the percentage and classes was examined. The results showed there were a large number of students who fell into the class of above average and very few who fell into average or good classes. The results

obtained from the classification from the fuzzy logic system differed. The performance values were more distributed on both sides of the average class.

The fuzzy logic expert system developed in these previous studies can be applied to teacher evaluation in several ways. Just as student-centered learning has the need to incorporate multiple measures for student assessment, reduce the degree of subjectivity in qualitative types of student assessment, and account for different levels of evaluation lenience, such is the case when assessing and evaluating teacher performance. The study of Jamsadekar and Mudholkar (2013) also provides a framework in order to compare the distribution of outputs in a fuzzy logic expert system with the distribution of outputs in a traditional method.

Review of Studies Using Fuzzy Logic Applied to Teacher Performance

Studies have addressed the need for fuzzy expert systems in teacher evaluation (Fourali, 1997; Cole & Persichitte, 2000; Djam & Mishra, 2013). However, there is little evidence-based research of fuzzy expert systems for effective teacher performance, and it remains an active area of research (Amin & Khan, 2009; Atta-ur-Rahman, 2013; Bhosale & Kamath, 2013; Chaudhari et al., 2012; Gupta & Dwahan, 2012; Khan et al., 2011; Kumar, 2013; Pavani, Gangadhar, & Gulhare, 2012; Ramli, 2009; Trstenjak & Donko, 2013). The purpose of almost all the articles reviewed in the literature was to show how fuzzy logic could be applied in multi-criteria decision-making. Empirical methods focused on the design of the fuzzy expert system and illustrations of its effectiveness were most often made by generalization, one numerical example, or the differences of teacher performance scores compared to the traditional approach in several cases

observed by a histogram (Bhosale & Kamath, 2013; Chaudhari et al., 2012; Gupta & Dwahan, 2012; Purnama-Dewi, Oka-Sudana, & Daarma-Putra, 2012; Trstenjak & Donko, 2013). The vast majority of the participants in the studies in the literature review were university or technical institutional faculty (Amin & Khan, 2009; Bhosale, & Kamath, 2013; Chaudhari et al., 2012; Gupta & Dwahan, 2012; Kumar, 2013; Trstenjak & Donko, 2013).

Chiang and Lin (1994) reported the first known fuzzy logic system to evaluate high school teachers' performance. Cheng, Wang, Tsai, and Huang (2004) incorporated questionnaires with fuzzy linguistic responses to appraise the performance of high school teachers. The researchers used the responses to weight five criteria in evaluating performance and used fuzzy logic arithmetic operations to create a ranking index to rank the teachers. The method of which the fuzzy logic arithmetic operations ranked the teachers was later reviewed and extended by Wang and Chen (2008). The results of both reports focused on the accuracy of the mathematical methods of the ranking index and did not make any claims to the overall effectiveness of the teacher evaluation system. In the study by Ramli (2009), the purpose was to improve Wang and Chen's (2008) approach by using an aggregation process, which used fuzzy linguistic variables when applying the fuzzy logic arithmetic operations.

Chaudhari et al. (2012) evaluated the performance of thirty instructors of a technical institution in India with the use of student feedback, student attendance, and academic development as the input of a fuzzy expert system. The researchers observed the difference in the direct value and the value determined by the fuzzy expert system, which was due to the weights given to some of the input variables when establishing the

fuzzy rules. The researchers claimed the observed difference provided evidence that the value from the fuzzy logic model was more realistic than the direct value. However, there was no evidence provided to support the claim that the value from the fuzzy logic model was more realistic.

Gupta and Dwahan (2012) developed a fuzzy logic system to measure the quantitative as well as qualitative aspects of university faculty evaluation in Punjab, India, with nine input variables classified as inside or outside class. The nine input vector consisted of five outside class parameters (library, forums, training and placement, research, and distance learning courses) and four inside class parameters (pedagogical initiative, presentation assessment, discipline, and assignment based test). The researchers claimed by examining the linguistic variables of both of these types of parameters, a fuzzy logic system could suggest improvements, but no empirical examples were conducted.

Bhosale and Kamath (2013) designed a fuzzy logic system to evaluate teachers using three modules: teaching and learning evaluation, professional development activities, and research and publication contributions. Using the input of ten selected cases, the researchers claimed that their fuzzy logic system was used to improve the efficiency of teaching staff performance. However, the purpose of both of these studies was to present how a fuzzy logic system can be used to build a performance evaluation model, and there was no evidence to support this claim.

Kumar (2013) developed a fuzzy logic system to examine the relationship between qualitative factors of teacher performance and external factors like compensation and incentives. Fifty questionnaires were distributed to teachers in engineering colleges

in India in order to determine the membership values. As a result, the fuzzy logic system, Performance-Incentive-Development (PID), was developed. This system adaptively changes the incentives based on the evaluation of teacher performance, but no cases were examined.

Amin and Khan (2009) and Khan et al. (2011) acknowledged a commonly cited problem in the design of a fuzzy logic expert system: how to efficiently and effectively acquire human expert knowledge and transform it into a suitable format for a computer during the knowledge acquisition phase. Amin and Khan (2009) developed a detailed knowledge acquisition process and a questionnaire as a knowledge collection tool. The responses of 25 faculty members in universities throughout Pakistan, when asked to rate the importance of each teacher performance criterion, were used to assign weights to the rules of the fuzzy logic expert system. The findings suggested that a questionnaire was an adequate tool for knowledge acquisition from multiple experts. Khan et al. (2011) utilized the results from the questionnaire in the development of the fuzzy logic expert system in their study. Three different examples were considered. The findings from the examples assigned a rating for teacher performance, but these ratings were not tested or compared to a traditional method.

In other studies, knowledge extracted during the knowledge acquisition phase is acquired through student feedback, in the form of questionnaires, to weight the measures of teacher performance (Atta-ur-Rahman, 2013; Chaudhari et al., 2012; Pavani et al., 2012; Ramli, 2009; Trstenjak & Donko, 2013). Atta-ur-Rahman (2013) proposed the Teachers Assessment and Profiling System (TAPS) as an approach to using a fuzzy rule based system to examine the weaknesses and strengths of university faculty through e-

questionnaires completed by university students. Atta-ur-Rahman (2013) claimed that this approach distinguished itself from others because, in addition to a fuzzy rule based system to extract information, a profiling Apriori Algorithm was also implemented to find associations and relationships between the criteria. The Apriori Algorithm used data of teacher performance over multiple semesters to find associations of teachers to class, subject, semester, and discipline. The research used an example to illustrate the results and effectiveness of the software. The findings were not empirically tested or compared to a traditional approach.

Similarly, Trstenjak and Donko (2013) developed a systematic assessment Performance Appraisal to evaluate teachers. A survey, given to 300 students enrolled in a computer engineering undergraduate degree program in a university in Croatia, asked students to evaluate the professor on criteria such as lecture quality, teacher-student relationships, approach to modernization of teaching, and others. The results were used to weight the rules in the fuzzy logic expert system. The method of the study used an illustrative example of three teachers. The researchers claimed from these results, by taking into consideration students' opinions and criteria with different weights, the model provided a more advanced and sensitive approach that could, ultimately, result in a more realistic evaluation.

Neogi, Mondal, and Mandal (2011) presented a methodology based on a cascaded fuzzy inference system to evaluate the performance of non-teaching staff of a university in west Bengal. Ten non-teaching staff were randomly selected and examined in a comparative analysis between the proposed cascaded fuzzy inference system and the

existing one. The study validated the design by calculating the percentage error in the fuzzified output with the statistical mean from the existing system as shown below.

$$\text{Percentage Error} = \frac{(\text{Fuzzy Logic Model Output} - \text{Mean Output})}{\text{Mean Output}} \times 100$$

The criterion for validation in the study by Neogi et al. (2011) was based on the condition that the values from the fuzzy logic model should not vary more than $\pm 10\%$ of the mean value.

In addition, the researchers conducted a sensitivity analysis to examine the accuracy of the model, and four different parameters were modified: membership functions, aggregation methods for both combining input variables and resulting rules, and defuzzification methods. The results claimed to acquire a better understanding of how membership functions and rule sets interact but stated that the process of empirically adjusting membership function parameters was a “very tedious and time-consuming task” (p. 610). Neogi et al. (2011) also claimed that modified aggregation or defuzzification methods provided very little change in the results.

Although there are studies, which have reported accuracy on teacher performance evaluation, there has not been a study conducted to date of fuzzy logic applied to teacher performance in the Western world or in the United States. In the forward written by Zadeh in Singh et al. (2013), he wrote, “for researchers in the engineering and scientific community, the word *fuzzy* is still *fuzzy*” (p. 1). It was suggested part of the reason fuzzy logic was not initially more widely accepted was because of the choice of the word *fuzzy* in the name. Djam and Mishra (2013) suggested another reason it has not gained wide acceptance could be due to the deficiency of these methods to provide the stakeholders with transparent and interpretable results.

The only research to date that uses fuzzy logic to measure any aspect of education policy in the United States is a study by Yates (2009), who developed a fuzzy logic expert system to determine if some Louisiana schools had been misclassified according to Adequate Yearly Progress (AYP) measures. His inquiry claimed that AYP assessment methods produced unstable results and other problems believed to produce poor validity. By using the Louisiana Adequate Yearly Progress (AYP) Plan, specifically the School Performance Scores star ranking system, twenty randomly selected schools in each “star” category were selected and evaluated by the fuzzy logic system. Yates (2009) reported no school with a passing AYP score was identified as having a high possibility of misclassification with the fuzzy logic system. The researcher did report examples of schools with a failing AYP score as having a moderate to high possibility of misclassification with the fuzzy logic system. The researcher claimed these examples indicated that the prototype successfully identified potentially misclassified schools (Yates, 2009).

Validity and Reliability of Fuzzy Logic Expert Systems

Although the true value of teacher performance is not known and cannot be directly measured, the validity of a fuzzy logic system for situations when the true value is known and can be measured provides evidence for a fuzzy logic approach. The validity of fuzzy logic has been considered in situations, such as in the medical sciences, where linguistic terms associated with variables can be directly measured (Ajiboye & Weir, 2005; Reiss, Hennessey, Rubin, Beach, Abrams, & Warsofsky, 1998). Reiss et al., (1998) proposed a fuzzy logic algorithm to accurately separate volume-based tissue

matter in the central nervous system, which was applied to high resolution, multispectral images. The validity and reliability of the fuzzy logic method was evaluated by assessing the stability of the algorithm across time, rater, and pulse sequence. The accuracy was applied to image datasets, and the differences in specific tissue volumes. In each case, the algorithm was found to have high reliability, accuracy, and validity (Reiss et al., 1998).

Another example where fuzzy logic has accurately measured what can be directly measured is in horticulture. The growth of a plant can be very uncertain and depends on many uncertain parameters such as shoot length, number of leaves, and root length of the plant (Mandal, Choudhury, De, & Chaudhuri, 2008). In Mandal et al. (2008), a fuzzy logic approach was used for the prediction of shoot length of a mustard plant at maturity. An error analysis was conducted between the predicted error and average error, and the results indicated fuzzification of data was appropriate to predict the shoot length.

However, the difficulty of comparing a fuzzy logic system to a traditional method for a social science subject is that the true value is not known, and situations with qualitative terms cannot be directly measured (Ganideh & Aljanaideh, 2013; Ganideh, Refae, & Aljanaideh, 2011; House, 2012). The reliability and validity of these identifications cannot be empirically tested because they are generated from human judgment, which is often neither reliable nor valid (Yates, 2009). To examine the results produced with a fuzzy logic approach, it is helpful to compare them to an established index (Al Ganideh, El Refae, & Aljanaideh, 2011; House, 2012).

In political science, difficulty comes in measuring qualitative or subjective attributes such as the level of democracy. House (2012) produced a fuzzy logic system

that specified the degree to which countries function as true democracies. To test and gauge the system's performance, the researcher examined the input values of eleven countries from diverse locations with diverse political systems. The outputs of the Democracy Index created by the fuzzy logic system were reported along with the Democracy Index created by *The Economist*. By examining the variable relationships, House (2012) suggested the fuzzy logic system was able to illuminate features of the system and even predict about countries.

Ganideh et al. (2011) used a fuzzy logic approach to predict the consumer ethnocentrism (CET) tendencies scale or CETSCALE scores, which expresses the level of ethnocentric tendencies a country's consumers show towards their national products. The correlation coefficients of CETSCALE were found reliable with each Cronbach's alpha of 0.70 or greater. The fuzzy logic system used the following inputs: patriotism, nationalism, and internationalism to determine the level of consumer ethnocentrism. The predicted CETSCALE scores by the fuzzy logic approach and the real measured CETSCALE scores obtained from 340 consumers' responses to a questionnaire distributed in market areas in Jordan were compared. The results indicated the fuzzy logic approach successfully predicted the score for that consumer as given by the questionnaire. Ganideh et al. (2011) argued traditional statistical techniques can only give insights to the nature and the strength of the relationships, but a fuzzy logic approach can help managers to accurately predict customer ethnocentric tendencies for a particular type of customer based on the input variables.

The study by Charles, Kumar, and Suggu (2013) proposed a fuzzy logic model to assess the service quality gap in the Malaysian banking sector because perceptions or

expectations of service quality are generally expressed in vague linguistic terms. The research constructed a fuzzy method using the SERVQUAL instrument, a survey of 26 items, to evaluate the service quality of banks in Malaysia for Islamic and foreign bank markets. Questionnaires are frequently designed with the Likert Scale to gauge the perception of the participants. Using the test of internal consistencies on dimensions of service quality, Charles et al. (2013) compared the effectiveness of the Fuzzy linguistic scale and the Likert scale. The ranges of reliability coefficient were between 0.697 and 0.860 for the Likert Scale and 0.908 and 0.957 for the Fuzzy linguistic scale. Thus, the report suggested that the Fuzzy linguistic scale created more reliability than the Likert scale, especially in the presence of a skewed distribution of customers' scores.

Among the important topics of reliability is how to evaluate a proper correlation coefficient with fuzzy data, especially when the data is uncertain or vague (Cheng & Chung, 2014). Generally, Pearson's correlation coefficient can be used to measure the correlation between two random variables. However, when the data are fuzzy interval values, it is not possible to use a classical approach to determine the correlation coefficient (Cheng & Yang, 2013; Cheng & Chung, 2014; Szmidt, Kacprzyk, & Bujnowski, 2012). Furthermore, the uncertainty in the statistical numerical data is the important point of the problem. Incorporating the uncertainty with traditional mathematical computation is difficult to establish and formulas in these studies are quite complicated, requiring mathematical programming (Cheng & Chung, 2014; Cheng & Yang, 2013).

Conclusion

Fuzzy logic has a unique applicability to the field of education, but there is much research to be done on the application of fuzzy logic to education and assessment. To date, there is no study of fuzzy logic expert systems used to assess teacher performance by new state teacher evaluation policy reform. As understanding of the complexity of human behavior and learning increases, the ways of describing and supporting that learning must increase as well. Research must extend beyond claims that fuzzy expert systems are more reliable and accurate than traditional methods with generalized or illustrative examples. This area of research must go beyond simply presenting the details of the fuzzy logic design in technical descriptions of the membership functions and present fuzzy logic systems in a way that allows stakeholders to understand the implications for education.

CHAPTER FOUR

METHODS

The research method that was used for this study was an empirical, exploratory case study. This method of inquiry is especially useful to test theoretical models by using them in real world situations and provides the basis for the application and extension of methods (Yin, 1994). In a continuing search to find the most precise methods for evaluating teacher performance, a fuzzy logic expert system designed to distinguish between levels of teacher performance was explored. A fuzzy logic expert system was designed using Matlab, a math software program, and the data from one Georgia high school. Human expert knowledge was extracted from four principal evaluators in order to write the rules that contributed to the design of the expert system. The summative scores of the principal evaluations of the participants volunteering their rating scores of the teacher performance standards were then calculated by the method currently used by the state and by the fuzzy logic expert system to see what additional, if any, insights the use of a fuzzy logic evaluation system could provide. To test the fuzzy logic expert system, two phases of software development, incorporating artificial data, were employed to ensure the overall validity of the fuzzy logic expert system.

Research Questions

The purpose of the study was to examine to what degree a fuzzy logic expert system could elicit and quantify teacher performance using state teacher evaluation methods. The study accomplished this by exploring the following research questions:

1. How can a fuzzy logic expert system quantify teacher performance using ratings from principal observations?
2. How do the ratings of the fuzzy logic expert system identify and distinguish levels of teacher performance as compared to traditional evaluation methods?

Participants

For this study, one unnamed Georgia public high school was used. The student body was approximately 1,050 students in grades nine through twelve. Historically, the school met Adequate Yearly Progress (AYP), the statewide accountability system. Under the new statewide accountability system, College and Career Ready Performance Index (CCRPI), each school receives a grade from 1 to 100. For 2013, the school's CCRPI (73.3) was higher than the state average (71.8), and for 2014, the school's CCRPI (64.6) was lower than the state average (68.4) (Georgia Department of Education, 2014b). The school had approximately 35 percent of students classified as having low socio-economic status, and the ethnicity is described as 85 percent Caucasian, 14 percent African-American, and one percent other. The school had one principal and three full-time assistant principals who conducted the TKES teacher performance evaluation.

The sixty-three certified teachers, who were evaluated under the TKES system, were asked to volunteer their rating score for this study through staff email and an announcement in a faculty meeting. Participation was anonymous, as no identifiers of

participants were given to the researcher. Of 63 certified teachers, 51 certified teachers (84%) participated in the study. Five did not respond to the request for participation, and five teachers declined to participate. Two were not present at the time of the study because of medical or personal leave and were not included in determining the participation rate.

Instrumentation

As part of the Race to the Top Initiative, Georgia piloted the Teacher Keys Effectiveness System (TKES) beginning in January of 2012. The TKES consists of the following two components: Teacher Assessment of Performance Standards (TAPS) (50%) and Student Growth Percentiles (50%). In order to obtain the TAPS measure, each teacher is observed two times for a minimum of 30 minutes. State policy recommends one observation be scheduled in advance and the other be unannounced; however, the decision is ultimately determined by the individual district. The school district where the study was conducted followed the state policy. These two 30-minute observations, known as formative assessments, were to inform one summative assessment of TAPS measure for the year. While the principal is ultimately responsible for all evaluation activities within the school, the school system may also designate assistant principals, department chairs, or any member inside or outside the school who has been trained as a TAPS evaluator for these observations. In the particular school chosen for this study, the principal and three assistant principals conducted the TAPS observations of teachers.

The Georgia Department of Education defines the TAPS observation approach as three-tiered. An example using the third standard, Instructional Strategies, can be seen in

Figure 12. The top tier consists of five domains or major categories in TAPS: Planning, Instructional Strategies, Assessment of and for Learning, Learning Environment, and Professionalism and Communication. The second tier consists of ten standards, grouped into five domains of two standards each. A complete list of these ten standards and their definitions can be found in Appendix A. The last tier is a list of performance indicators, which are examples of the types of observable behaviors that may occur for each standard. It is worthy to note that the list of performance indicators is not exhaustive and is not intended to be used as a checklist but serve as a guide (Georgia Department of Education, 2013b).

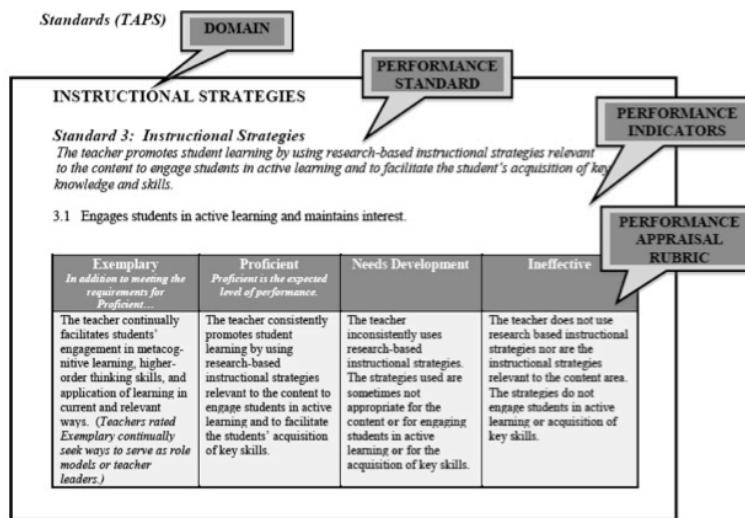


Figure 12. Example using showing the relationship between the three tiers of the TAPS. Reprinted from Georgia Department of Education (2013b). Copyright 2013 by Georgia Department of Education. Reprinted with permission.

During formative assessments, principals use performance appraisal rubrics to rate teachers on each of the ten standards of the Teacher Assessment of Performance Standards (TAPS). The performance appraisal rubric provides a qualitative explanation

of each of the four levels of performance: *Ineffective*, *Needs Development*, *Proficient*, and *Exemplary*. The definition and description of these ratings can be seen in Appendix B. The description provided in the *Proficient* level is the expected level of performance. Teachers who earn an *Exemplary* rating must not only meet the requirements for the *Proficient* level but also exceed them (Georgia Department of Education, 2013b).

In calculating the overall score of each teacher in the summative assessment, each standard is rated and assigned as follows: *Exemplary* ratings are worth 3 points, *Proficient* ratings are worth 2 points, and *Needs Development* ratings are worth 1 point. *Ineffective* ratings have no point value. The summation to establish the overall score is used to determine the category of performance of the teacher. An example calculation can be seen in Table 1, and the summative point values for each category of performance rating can be seen in Table 2. In the example, the total value of all ratings (19) falls in the *Proficient* category, so this teacher would be classified as *Proficient*.

Table 1

Example of Overall Summative Rating

Rating/Overall Point Value	Point Value	Number of Standards	Computation
Exemplary	3	2	$3 \times 2 = 6$ points
Proficient	2	6	$2 \times 6 = 12$ points
Needs Development	1	1	$1 \times 1 = 1$ point
Ineffective	0	1	$0 \times 1 = 0$ points
			Total = 19 points

Table 2

TAPS Final Ratings from Summative Scores

Final Ratings	TAPS Summative Cut Scores
Ineffective	0-6

Final Ratings	TAPS Summative Cut Scores
Needs Development	7-16
Proficient	17-26
Exemplary	27-30

Validity and Reliability of the Teacher Assessment of Performance Standards

Reliability and validity are important to provide a robust research design in any study. If an evaluation system cannot validly measure teacher performance, then it cannot be used in high-stakes decisions, and “when educators realize this, they will be less motivated to perform” (Milanowski, 2007, p. 4). The criterion-related validity of the observation instrument used, or whether the instrument actually measures what it claims, is dependent on the correlation of the value-added measures of teacher performance and student achievement scores (Barnett et al., 2014; Corcoran, 2010; Measures of Effective Teaching 2010; Rockoff et al., 2011; Rockoff & Speroni, 2010).

Even with the whirlwind of activity around teacher evaluation since 2009, the reality of the availability of data is that most states have barely begun to implement these new systems (National Council of Teacher Quality, 2013). Furthermore, although a vast amount of research on value-added models exists, only a minority of teachers is subject to an evaluation based on the test gains of students (Whitehurst et al., 2014). Because the data of the student growth and achievement scores are still being collected, it is not yet possible to examine the relationship between teacher performance and student achievement measures, and many of the validation questions cannot be answered (Stronge et al., 2013). Presently in Georgia, only teachers who teach courses that

administer the criterion-referenced End of Course Test (EOCT) have student achievement scores. In the pilot year 2014-2015, the student achievement scores were not used or available to calculate into the overall Teacher Effectiveness Measure. For this reason, the study exclusively used principal evaluations of teacher performance with the TAPS standards as the teacher performance measure, and a value-added score was not used to measure the validity of the TAPS instrument.

Validity measures of the standards outlined in CLASS KeysSM, which was selected as the framework for the TAPS standards and primarily developed by Stronge & Associates, (Georgia Department of Education, 2011; Stronge & Tonneson, 2011) have been reported. Validity measures of Stronge's Teacher/Leader Effectiveness Performance Evaluation System, as adopted by the Department of Education in Virginia and in other studies, produced consistently significant reliability measures (Stronge et al., 2013; Stronge, Ward, & Grant, 2011; Stronge, Ward, Tucker, & Hindman, 2008).

Stronge et al. (2013) reported higher validity measures between composite scores of teacher performance standards and the mean student achievement gains of the Stronge Evaluation System when compared to the findings of Cincinnati Public Schools, which implemented Danielson's (1996) Framework for Learning (Gallagher, 2004). Correlation values with the Stronge teacher evaluation system ranged from 0.28 to 0.38, and all correlation values for Cincinnati Public Schools were lower than 0.30 (Gallagher, 2004; Stronge et al., 2013). These correlation values are not statistically significant, consistent with the controversy of the validity of value-added measures as an appropriate tool for identifying effective teachers (Goldhaber & Loeb, 2013; Hallinger et al., 2014; Kane et al., 2008; Schochet & Chiang, 2010).

The Department of Education in Virginia adopted the Stronge Teacher

Performance Evaluation system for its Teacher Performance-pay Initiatives 2011-2012 pilot (RMC Research Corporation, 2013). A validation study by Stronge et al. (2013) included statistical analyses by examining the relationship between the ratings of over 300 teachers on seven performance standards. The rating of the Professional Knowledge standard had the most overlap with other standards, meaning it was significantly correlated with the Instructional Delivery, Assessment, and Professionalism standards. The Learning Environment and Professionalism standards were the most unique, meaning the lowest level of correlation, $r = .472$ (Stronge et al., 2013). With the validity of the TAPS instrument established in the study by Stronge et al. (2013), further validity measures were not calculated in this study.

There are several types of reliability of interest when examining the TAPS evaluation. Each of these types is important for establishing a reliable evaluation system and is discussed below. Criterion-related reliability refers to the evaluators' level of training and the extent to which a trained evaluator's ratings agree with those of an expert evaluator. This type of reliability is important because it assures that the level of understanding of the trained evaluator is the same as that of an expert (Strong et al., 2013). The Georgia Department of Education (2013b) acknowledges the need for evaluator training and credentialing and requires evaluators to participate in TAPS training and successfully complete the Evaluator Credentialing Assessment. The evaluator credentialing ensures that an evaluator has the minimum qualifications to evaluate the teacher with the TAPS rubric. Criterion-related reliability, as well as inter-

rater reliability discussed below, is typically established by training, and ongoing professional learning is necessary to maintain and deepen the level of reliability (Daley & Kim, 2010; Georgia Department of Education, 2013a; Stronge et al., 2013).

A second type of reliability of interest is the intra-rater reliability, which refers to the extent to which the evaluators are consistent in their own ratings. Because this study examined one formative assessment, it was not necessary to calculate the intra-rater reliability. The third type of reliability, inter-rater reliability, assures two or more evaluators demonstrate the same or similar ratings.

In 2009, faculty members at Valdosta State University, working with Georgia Department of Education, conducted the pilot reliability study of the CLASS KeysSM teacher observation instrument. Volunteer administrators were asked to rate four fifty-minute, digital recordings of teachers delivering instruction to students. The reliability data across teachers, occasions, and raters was averaged to yield an overall reliability measure. Using the percentage of absolute agreement, the researchers claimed that reliability measures were significant, but the exact values were not reported (Georgia Department of Education, 2011).

Estimates of inter-rater reliability are most commonly measured by the intraclass correlation coefficient (ICC) (Graham, Milanowski, & Miller, 2012). An ICC score of 1 indicates perfect agreement while a score of 0 indicates no agreement. In general, there is consensus that an ICC value of .70 is sufficient for research purposes, but others advocate a value of .8 as a minimum when using scores for high-stakes purposes such as compensation, retention, or promotion (Graham et al., 2012). This study proposed to calculate the inter-rater reliability among the four evaluators using the intraclass

correlation coefficient (ICC) with a two-way random effects ANOVA model (Shrout & Fleiss, 1979).

Design of the Fuzzy Logic Expert System

The first research question asked, “How can a fuzzy logic expert system quantify teacher performance using ratings from principal observations?” In order to explore this question, a fuzzy logic expert system was developed. In the design of the fuzzy logic expert system, there were many decisions that had to be made. The designer chose the method for extracting knowledge from experts, the type of software, the type of fuzzification, the number and form of membership functions, the parameters of the membership functions and their operator, the implication and aggregation operator, and the type of defuzzification. This demonstrated the richness of fuzzy controllers but also the need for some guidelines for their practical design (Bourchon-Meunier, Dotoli, & Maione, 1996). The following section lists the decisions that were made in each step of the design process and the rationale for these choices.

Knowledge Acquisition Process

The system design followed the same process as the design of a fuzzy logic expert system previously done in many recent works (Hong & Lee, 1996). In Figure 13, one sees that the researcher or knowledge engineer had dialogue with the human experts in the domain, and the explicit knowledge acquired was used in the expert system knowledge base. The facts and expertise can then be transformed to any user of the system.

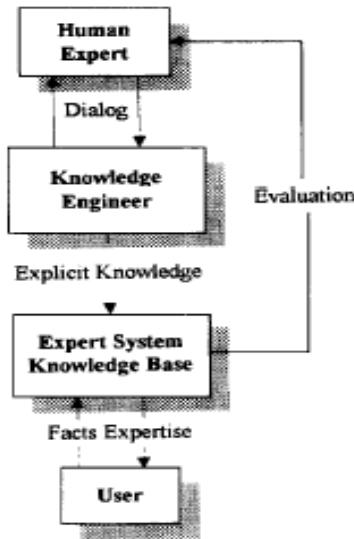


Figure 13. Development of a classical expert system (Hong & Lee, 1996).

For this study, the first phase of the construction of the fuzzy logic expert system began with acquiring qualitative human knowledge from experts in the field of education. This process known as knowledge acquisition was conducted through the use of a questionnaire. Similar studies have used questionnaires in the knowledge extraction phase (Cheng et al., 2004; Amin & Khan, 2009; Khan et al., 2011; Kumar, 2013; Wang & Chen, 2008). This type of approach is recommended to prevent the “bottleneck” that can occur in research design (Amin & Khan, 2009; Khan et al., 2011). The questionnaire was distributed through email to the four administrators of the selected high school requesting a reply to the email within two weeks. The questionnaire asked the principals to rank the 10 standards from the TAPS in order of importance and influence on teachers’ performance. Rankings were from 10 (*most* critical to teacher performance) to 1 (*least* critical to teacher performance). The complete questionnaire can be seen in Appendix A. For each standard, the initial results from the four administrators were averaged and then weighted by the rank sum method (Edwards & Newman, 1982) as can be seen in Table 3.

In this method, each mean is normalized by dividing by the sum of all ranks. This assures that the sum of the weights for the ten standards is one. These weights were then used in the design of the fuzzy logic expert system to weight the rules, which is further discussed in the second step of the design.

Table 3

Response Summary for Effect on Teachers' Performance

TEACHER ASSESSMENT OF PERFORMANCE STANDARDS (TAPS)	Mean	Weight
1 Professional Knowledge	8	.1455
2 Instructional Planning	5.75	.1045
3 Instructional Strategies	8	.1455
4 Differentiated Instruction	3.25	.0591
5 Assessment Strategies	5.75	.1045
6 Assessment Uses	3.25	.0591
7 Positive Learning Environment	8.5	.1545
8 Academically Challenging Environment	4.75	.0864
9 Professionalism	3.5	.0636
10 Communication	4.25	.0773

In addition to ranking each of the ten standards, the knowledge engineer, or researcher, met for one hour with the four administrators on December 18, 2014, to assist in the preliminary development of the rules. This meeting was recorded with voice memo on an electronic device. The recording was housed in a locked cabinet until it was transcribed to a password-protected computer file, at which point the voice memo was erased. The result of the meeting gave additional insight or knowledge, which was extracted to form additional rules to the 80 rules later discussed in the second step of the design of the fuzzy logic expert system. The administrators were asked to answer the following questions:

- 1) In comparing instructional planning and strategies with classroom environment, do you agree or disagree that a teacher who is rated higher in classroom management is overall a more effective teacher than one who is rated higher in

classroom instruction?

2) Do you agree or disagree that the standards comprising the TAPS should be equally weighted?

3) Are there any two standards that predict each other? Are there any standards that measure the same thing?

The answers from the administrators are discussed in the second step of the design of the fuzzy logic expert system when assigning rules.

The software used to design the expert system was the Fuzzy Logic Toolbox 2.2.7 in Matlab by Mathworks, Inc. (<http://www.mathworks.products/fuzzylogic>). This software was chosen because of its ease in understanding and common use among the fuzzy logic knowledge engineering community. A Mamdani-type fuzzy inference system was used in this study, as it is the one most commonly used to reflect human-based rules (Mathworks, Inc., 2013; Neogi et al., 2011). The four steps in constructing the fuzzy logic expert system in Matlab are described below.

Step 1: Fuzzification of the Input Variables

The ten TAPS performance standards were the ten input variables listed in Table 4. The input range of values (1-4) was identical to the categorical rating given by the evaluators by the TAPS performance appraisal rubric, and the membership functions were defined as *Ineffective*, *Needs Development*, *Proficient*, and *Exemplary*. It is important to note that while the administrators used the scale 1-4 to evaluate the levels of performance, the point values assigned to these levels when determining the summative score were 0-3 as defined by the Georgia Department of Education.

Table 4*Input and Output Variables Defined*

Input Variables	Output Variable
1. Professional Knowledge 2. Instructional Planning 3. Instructional Strategies 4. Differentiated Instruction 5. Assessment Strategies 6. Assessment Uses 7. Positive Learning Environment 8. Academically Challenging Environment 9. Professionalism 10. Communication	Teacher performance

The fuzzification step plays an important role in a fuzzy environment, and this step is based on the membership function. The selection of a particular membership function depends on the nature of data to be used. Studies have highlighted the importance of the correct selection of a membership function (Garibaldi & John, 2003; Zhao & Bose, 2002). Thus, caution is also needed in the selection of a membership function to a given situation. If not, the resulting output will be improper and have high error (Mandal et al., 2008). Several differences in membership selection have been investigated in order to design a stable and reliable system. Psychological and statistical methods have been used in other studies in the choice of membership functions (Bouchon-Meunier, et. al, 1996; Ahmad & Asri, 2013).

For this study, the degree to which the input belongs to each of the appropriate fuzzy sets was represented with a triangle membership function, commonly used because of its simplicity and easy computations (Saleh & Kim, 2009). A fuzzy variable and its membership functions are defined and shown in Figure 14.

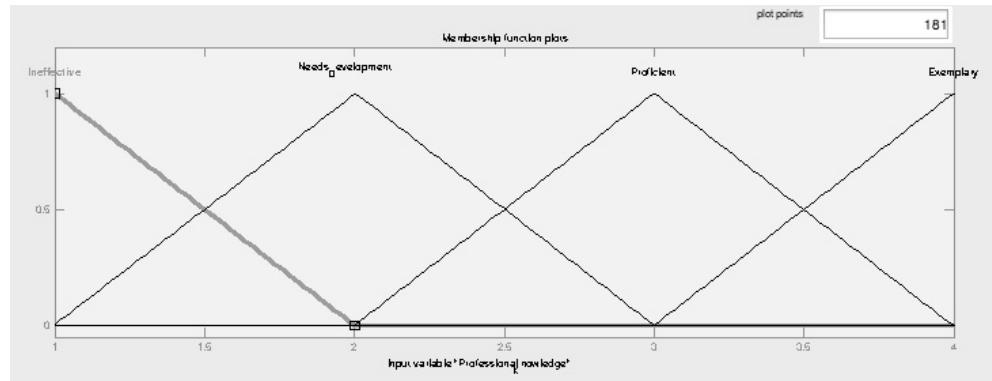


Figure 14. Example of the membership functions of the input variable, Professional Knowledge. The graphs display the degree in which it belongs to the four fuzzy sets: *Ineffective*, *Needs Development*, *Proficient*, and *Exemplary*.

The input data were the ratings assigned during the participants' TAPS evaluation and had complete or no membership belonging to the set. For example, looking at Figure 14, there are four values in the TAPS rubric that correspond to four categories (*Ineffective*, *Needs Development*, *Proficient*, and *Exemplary*). A rating of 2 reflects a total degree of membership (1) in the category *Needs Development*. If a rating of 2.5 could be assigned to teacher performance, the value would have a 0.5-degree membership in both categories, *Needs Development* and *Proficient*. Since the range of values of the traditional evaluation system did not allow for membership as a matter of degree, the simplicity of a triangle membership was appropriate.

Step 2: Assigning Rules

The data collected from the questionnaire to the administrators was used to assign weighted numerical levels of importance to the performance standards from the perspective of the individual education community. Fuzzy rules were then assigned to reflect the qualitative nature of human decision-making. Decision-making matrices have

been created for making decisions between the TAPS ratings of teacher performance and student growth measures, in order to categorize the overall teacher effectiveness measure (National Council on Teacher Quality, 2013; Georgia Department of Education, 2013b).

This study built the knowledge base of rules for the fuzzy logic expert system by a similarly designed decision-making matrix as shown in Figure 15.

Professional Knowledge	Exemplary	Needs Development	Proficient	Exemplary	Exemplary
	Proficient	Needs Development	Proficient	Proficient	Exemplary
	Needs Development	Ineffective	Needs Development	Proficient	Proficient
	Ineffective	Ineffective	Ineffective	Needs Development	Needs Development
Ineffective		Needs Development		Proficient	
Instructional Planning					

Figure 15. Example of a decision matrix. The matrix displays the rating for a domain based on the rating of two individual standards.

In the first step, the input variables were fuzzified, and values that represent the degree to which they belong to the corresponding fuzzy sets in the rule antecedent were listed. Before applying the implication method, the antecedent of each rule was assigned a weight according to the level of criticalness obtained from the responses to the questionnaire from the four administrators. The fuzzy operator “*and*” was used to formulate the conditional statement. The value of the consequent of the rule was a membership value in the output fuzzy set. For example, consider the rule, “If professional knowledge is *Exemplary* and instructional planning is *Ineffective*, then teacher performance is *Needs Development*.” Suppose the weight of the input variable, Professional Knowledge, was 0.2, and the weight of the input variable, Instructional Planning was 0.3. Because the antecedent of the rule was joined by the fuzzy operator

“and,” the weight assigned to the rule in the fuzzy logic expert system was the minimum value (0.2). Table 5 shows the four conditional statements or rules that were extracted from the first row of the decision-making matrix in Figure 15.

Table 5

Subset of Rules Extracted from the Decision-making Matrix.

If professional knowledge is exemplary and instructional planning is ineffective then teacher performance is needs development.
If professional knowledge is exemplary and instructional planning is needs development then teacher performance is proficient.
If professional knowledge is exemplary and instructional planning is proficient then teacher performance is exemplary.
If professional knowledge is exemplary and instructional planning is exemplary then teacher performance is exemplary.

Thus, a total of sixteen rules were written for the entire matrix. Each of the four other domains or pairs of the TAPS standards were equally represented as a design-making matrix. Only standards in the same domain were paired since the domains define standards, which can be paired together. Each domain resulted in 16 rules yielding a total of 80 rules. In addition to these rules, more rules were created as a result of the discussion with the administrators. The following summarizes the consensus of the answers to the following three questions asked by the researcher.

The reports of Kane et al.’s (2011) study provided the rationale for the first question the researcher asked the administrators: In comparing instructional planning and strategies with classroom environment, do you agree or disagree that a teacher who is rated higher in classroom management is overall a more effective teacher than one who is rated higher in classroom instruction? Kane et al. (2011) reported that for students

assigned to different teachers with the same overall classroom observation score, math achievement was higher for students whose teacher was rated higher in classroom management than students whose teacher was rated higher in instructional practices. Although classroom management is not a standard of TAPS, classroom environment is a domain of TAPS. In response to the first question regarding instructional strategies and classroom environment, the administrators agreed, like many of the TAPS standards, there is a connection. Identifying the exact relationship between the standard, Instructional Strategies, and the domain, Learning Environment, is dependent on many factors including the characteristics of the individual teacher. For some teachers, “sound instructional strategies lead to less management problems in the first place” (administrator interview, December 18, 2014). While for other teachers, “classroom environment may support all the rest or make it all fall apart” (administrator interview, December 18, 2014). From this discussion, the researcher created additional rules pairing Standard 3: Instructional Strategies with both Standard 7: Positive Classroom Environment and Standard 8: Academically Challenging Environment. The rules were formed using the previous decision-making matrix. Thus, these two additional pairing of standards produced sixteen rules each and 32 additional rules.

The responses varied regarding the second question: Do you agree or disagree that the TAPS standards should be equally weighted? While there was somewhat of a consensus that some standards were more critical than others, they agreed equal weighting prevents teachers from focusing too much on any one particular standard. With this information, the researcher decided to weight the rules according to the responses of the evaluators on the questionnaire listed in Table 3 but observed that the

values calculated did not heavily weight any one standard.

The third question asked by the researcher was the following: Are there any two standards that predict each other or measure the same thing? In response to the third question, the administrators collectively agreed there was also a connection between Standard 4: Differentiation and both Standard 5: Assessment Strategies and Standard 6: Assessment Uses, mainly “assessment strategies and uses are the means in which differentiation can be achieved” (administrator interview, December 18, 2014). These two additional pairings created 32 additional rules, similarly to the rules created from the responses to the first question the researcher asked the administrators.

One of the issues in assigning rules for a fuzzy logic expert system is referred to as the “curse of dimensionality” (Lee, Chung, & Yu, 2003). This occurs because the complexity of a problem increases exponentially with the number of variables as shown in the following equation: $N = p^m$, where p is the number of linguistic terms for each input variable and m is the number of input variables (Lee et al., 2003). In this study, designed for ten input variables and four linguistic ratings for each, the total number (N) of possible rules is $N = 4^{10} = 1,048,576$. Methods to reduce the total number of involved rules and the “curse of dimensionality” have been created by modeling of hierarchical fuzzy systems, where the total number of rules only increases linearly instead of exponentially (Lee et al., 2003). However, the number of rules does not correlate to the effectiveness of the system, and for this study, one did not need to write all possible rules. For example, if there was not a rationale provided by previous studies, the researcher, or the administrators to write a rule for professional knowledge and communication, then it was not necessary to write a rule with these input variables.

Steps 3 and 4: Aggregation and Defuzzification of the Output Variables

The process of aggregation is the process by which the fuzzy sets that represent the outputs of each rule are combined. If the input value has membership in a fuzzy set according to the corresponding membership function, then any rule with a non-zero output membership value is said to “fire” or produce a result. On the contrary, for example, if a teacher receives a rating whose degree of membership in the output set *Ineffective* is 0, the rules concerning *Ineffective* will not “fire.” The truncated output fuzzy sets for each rule are then aggregated into a single fuzzy region using the Mamdani model of truncated membership functions. Lastly, the fuzzy region is defuzzified, or converted into a crisp output value of teacher performance, by the centroid or center of gravity method. The formula for the centroid of a polygon is given below.

A polygon defined by n vertices $(x_0, y_0)(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$ is the point (C_x, C_y) , where $C_x = \frac{1}{6A} \sum_{i=0}^{n-1} (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$ and $C_y = \frac{1}{6A} \sum_{i=0}^{n-1} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$, where A is the polygon’s decomposed area, $A = \frac{1}{2} \sum_{i=0}^{n-1} (x_i y_{i+1} - x_{i+1} y_i)$ (Bourke, 1997).

The range of the output variable (0-30) was also the same as those outlined in the final ratings of the TAPS summative cut scores as previously shown in Table 2. The triangle membership functions, *Ineffective*, *Needs Development*, *Proficient*, and *Exemplary* are given in Figure 16, and the parameters given to the membership functions, as seen in Table 6, reflect the ranges of the summative cut scores determined by the state. The parameters for the membership functions *Needs Development* and *Proficient* were determined by the researcher with the logic that total membership in each of the ratings would occur in the middle of the range of the summative scores. For example, total

membership in *Needs Development* occurs at 11.5, the middle of the range 7-16. As this number increases, so does the membership in the set of the higher category. When this number has increased to 16, the upper limit of the range, there is equal membership of 0.5 in both the set of *Needs Development* and *Proficient*.

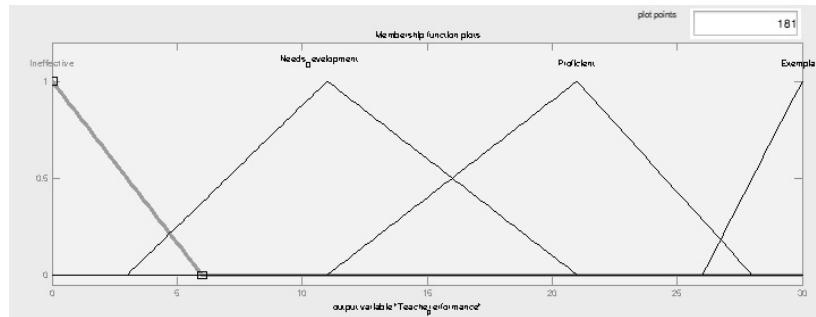


Figure 16. The membership functions of the output variable fuzzy sets. The graphs display the degree to which it belongs to the fuzzy set.

Table 6

Parameters of the Membership Function of the Output Variable Teacher Performance

Linguistic Variable	Parameters
Ineffective	[-2.5, 0, 6]
Needs Development	[0, 11.5, 21]
Proficient	[11, 21.5, 30]
Exemplary	[26, 30, 32.5]

With this value, a category rating of teacher performance can be assigned, as displayed in Table 7, that is consistent with the TAPS range of scores for each category. The output values for a fuzzy logic expert system are continuous values, whereas the values for the traditional method are discrete. Thus, the table was adjusted accordingly with the conventional rounding method. If the output value of the fuzzy logic expert

system is 26.5, the value could be rounded to 27, which would classify it as *Exemplary*, consistent with the traditional method.

Lastly, an overview of the design of the fuzzy logic expert system can be seen in Figure 17. The figure displays a portion of the input variables and rules while defining each of the steps in the fuzzy logic expert system.

Table 7

Mapping Between Crisp Output Ranges of the Fuzzy Logic Expert System and Linguistic Performance Values

Fuzzy Expert System Output	Linguistic Description
$26.5 \leq X \leq 30$	Exemplary
$16.5 \leq X < 26.5$	Proficient
$6.5 \leq X < 16.5$	Needs Development
$0 \leq X < 6.5$	Ineffective

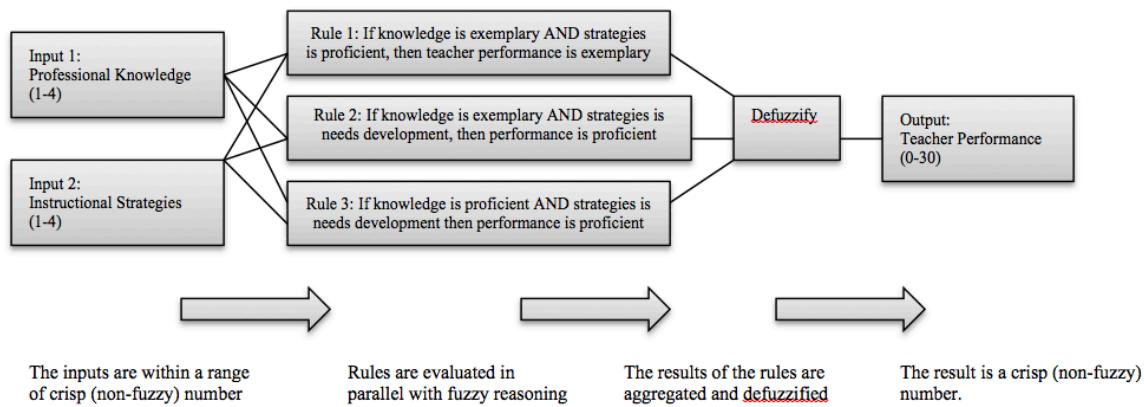


Figure 17. Portion of the design of the fuzzy logic expert system.

Testing and Validating the Fuzzy Logic Expert System

In order to validate the fuzzy logic expert system, the results were tested in a similar manner to those outlined in Yates' (2009) dissertation. In the study, Yates

(2009) used two common phases of a software development project: the proof of concept phase and the prototyping phase. A proof of concept is typically the first phase of a software development project that uses artificial data to test the overall design so that technical issues can be solved before moving on to the prototyping phase (Yates, 2009). Analysis of the results of proof of concept system is often informal.

Although the true value of teacher performance is not known and cannot be directly measured, the validity of a fuzzy logic system for situations when the true value is known and can be measured provides evidence for the validity or otherwise of a fuzzy logic approach. A common technique in software development during the proof of concept phase incorporates markers, or artificial data with these known values, into the system as a quality monitor (Yates, 2009). This study used the marker of a teacher who received all *Exemplary* ratings. If the system did not return the highest output, this would have indicated that there was a problem with the validity of the fuzzy logic expert system. Similarly, the marker of a teacher who received all *Ineffective* ratings should have yielded the lowest output of teacher performance. Additionally, a marker for a teacher with all *Proficient* ratings was used to compare the output of the fuzzy logic expert system to the output of the traditional method (20). This study used these markers to resolve technical issues and made necessary adjustments before moving to the prototyping phase. While there have been efforts to develop a reliability coefficient for fuzzy logic methods (Cheng & Chung, 2014; Cheng & Yang, 2013; Szmidt et al., 2012), these statistical methods are beyond the scope of this study and were not examined.

The results for this study were also analyzed informally by examining the outputs in the surface views given for the fuzzy logic expert system in Matlab. Surface views are

useful to show relationships and find optimum combinations between large amounts of data that may otherwise be difficult to see. Two input variables are plotted on the x and y-axis, and the output is plotted on the z-axis. From the surface view, one can quickly see the output value for any possible combination of input variables.

The second research question was, “How do the ratings of the fuzzy logic expert system identify and distinguish levels of teacher performance as compared to traditional evaluation methods?” This question was answered during the second phase of software development referred to as the prototype phase, which begins to simulate the full system or at least a part of it. The prototype phase is normally used to answer questions about the system (Yates, 2009). This was done in the literature reviewed exclusively by providing an empirical example of several scenarios. This study incorporated both the experimental results of the 51 participants from the selected school in the case study and the artificial data that simulated part of the system in order to explore the second research question. This study modeled the results of the artificial data of the prototype phase similar to the approach by Jamsandekar and Mudholkar (2013). Results of the outputs in both the traditional method and the fuzzy methods were plotted and labeled according to the categories of performance. Histograms of the frequencies of outputs in each category for both methods are then compared.

Artificial Data

Exemplary Range

The context of the artificial data first considered every possible combination of standards, which under the traditional method yielded a summative score of *Exemplary*

(27-30). When considering a combination, the ten ratings of the individual standards are referred to collectively as a data point and are represented, for example, as [3 3 3 3 3 3 3 3 3] with the first value representing the rating of Standard 1, the second value representing Standard 2, and so forth.

When considering all possible combinations of data points, only data points that were combinations of scores of 3 or 4 were considered. Consequently, there are 175 possible combinations of data points consisting of scores of 3 or 4, which would classify as *Exemplary* under the traditional method. This is found by the formula, $\sum_{i=0}^3 C_i$, where i is the number of scores of 3. The number of possible combinations, when considering a combination of three scores, such as 2, 3 or 4, grows exponentially, and for this reason, was not considered. It is noted that in order to maintain an *Exemplary* rating, or summative rating of 27, the maximum number of scores of 3 in the data point would be three, such as [4 4 4 4 4 4 4 3 3 3]. It is also important to note while the school used the scores 1 – 4 to rate the levels of performance, in order to obtain the point-values in order to calculate the summative score, 0 – 3 must be used. For example, the summative score for this data point is calculated as $3 + 3 + 3 + 3 + 3 + 3 + 3 + 2 + 2 + 2 = 27$.

Upper and Lower Range of *Proficient*

Secondly, the artificial data used every possible combination of standards that under the traditional method would be the lower and upper range of summative scores, which would classify as *Proficient* (17-26). The researcher defined the lower range of summative scores as the lower third of values (17-20) and the upper range as the upper third of values (23-26). In order to examine the lower range, only combinations of scores of 2 and 3 were considered by the same rationale as previously described, that the number

of possible combinations grows exponentially when considering all possible combinations of more than two different scores. Similarly, there are 175 possible combinations of data points consisting of scores of 2 or 3, which would classify as *Proficient* under the traditional method. This is found by the formula previously mentioned, $\sum_{i=0}^3 C_i$, where i represents the number of scores of 2. When examining the data points for a summative score close to the minimum value (17) of an overall rating of *Proficient*, it is noted that only three ratings of *Needs Development*, or scores of 2, are possible. For example, ratings of the ten standards could be [2 2 2 3 3 3 3 3 3] and still be classified as *Proficient*.

In order to examine the upper range of *Proficient*, combinations of scores of 3 and 4, whose summative scores ranged from 23-26 were considered. When examining the data points close to the maximum value (26) of an overall rating of *Proficient*, it is noted that six *Exemplary* ratings, or scores of 4, in a data point, such as [4 4 4 4 4 4 3 3 3 3], are possible.

Conclusion

In a continuing search of finding the most precise methods for evaluating teacher performance, a fuzzy logic expert system was designed to distinguish between levels of teacher performance. The scores of participants from one Georgia high school on ten teacher performance standards were the input values that were fuzzified by the membership functions in the first step of the fuzzy logic expert system. The rules of the system were written in accordance to decision-making matrices from the state, and additional rules were extracted and weighted from a survey and interview with the

principal evaluators. These rules were evaluated in parallel during the second step of the design. Lastly, the rules were aggregated and defuzzified into a crisp output of teacher performance. These values were compared to the method currently used by the state to see what additional, if any, insights the use of a fuzzy logic evaluation system could provide. In order to test the fuzzy logic expert system, two phases of software development, which also incorporated artificial data, were employed to ensure the overall validity of the fuzzy logic expert system.

CHAPTER FIVE

RESULTS

The purpose of this study was to examine to what degree a fuzzy logic expert system could elicit and quantify teacher performance using state teacher evaluation methods. In order to accomplish this goal, the following research questions were explored: How can a fuzzy logic expert system quantify teacher performance using the ratings from principal evaluations? How do the ratings of a fuzzy logic expert system identify and distinguish levels of teacher performance as compared to traditional state evaluation methods?

The methodology used to explore these research questions was an empirical exploratory case study. In order to answer the first research question, a fuzzy logic expert system was designed. The rules were weighted according to the survey given to the administrators. Additional rules were added to the rule base as a result of the interview with the administrators. During the proof of concept stage, adjustments were made to the fuzzy logic expert system. In order to answer the second research question, how do the ratings of a fuzzy logic expert system identify and distinguish levels of teacher performance as compared to traditional state evaluation methods, the certified teachers of an unnamed high school were asked to volunteer their ratings through staff email and an announcement in a faculty meeting. The four administrators rated teachers on the ten standards during one formative assessment.

In order to test the overall validity of the system, the first phase of software development, proof of concept, was implemented by using markers or data points where the true value is known and by examining the surface views. The data from the selected high school for the case study were compared to the output given by the fuzzy logic expert system. In order to further compare the methods, the artificial data provided in the prototyping phase was also further analyzed.

Research Question 1

Lack of a Perfect or Null Score Rating

In order to explore this question during the proof of concept stage, the researcher incorporated markers, or artificial data with known values, into the system as a quality monitor. During the proof of concept stage, there were several revisions made to the fuzzy logic system. The most notable was the adjustment of the parameters of the *Exemplary* and *Ineffective* membership functions in the output variable, Teacher Performance. This was decided based on the fuzzy logic expert system's output for the marker or data point, which contained scores of 4 for every standard. It is important to note while the school used the scores 1 – 4 to rate the levels of performance, in order to obtain the point-values and calculate the summative score, 0 – 3 had to be used. The summative score for this data point containing all 4's was calculated as $3 \times 10 = 30$. Logically, this output should be the same as the traditional approach, the maximum summative score of 30. Similarly, the output of the marker or data point consisting of scores of 1 for every standard should yield a summative score of 0.

The parameters for the *Exemplary* and *Ineffective* membership functions were adjusted so that the outputs of the data points for both extremes were close to the desired effect, 29.2 and 2.96. The fact that these numbers were only close to the desired effect was due to the defuzzification method, which for this fuzzy logic expert system, was the centroid method. Visually represented, the centroid finds the balancing point or center of the area of the region, which would not occur at the right boundary or maximum score of 30. In the same manner, the centroid would not occur at the left boundary or minimum score of 0. Other choices of defuzzification methods such as largest of maximum, which is based on the maximum value assumed by the aggregated membership function, were also explored but did not increase the maximum value. Lastly, the third marker was the data point of all *Proficient* ratings. The output of the fuzzy logic expert system was 19.6, a value sufficiently close to the target value of 20 assigned by the traditional method.

Surface Views

When examining the surface views of the fuzzy logic expert system, it is important to note that only two input variables may be compared simultaneously on the x and y-axis while the one output variable, teacher performance, is plotted on the z-axis. The surface views explained and listed in Figures 20 – 24 show all possible combinations of two input variables when the other eight input variables were fixed to 3, the most common rating. The fuzzy logic expert system contained 45 different pairings of two input variables. Each input variable pairing can be categorized into one of three types: input variables of the same domain of TAPS, input variables of different domains of

TAPS, which were paired when the additional rules were written, and the input variables of a different domain of TAPS not paired with the additional rules written.

The results of the surface views cannot be generalized easily. However, each of 45 different pairings depicts the general shape of one of three surface views listed below. The first surface view, Figure 18, depicts the input variables Professional Knowledge and Assessment Strategies, an example of a surface view of two input variables of a different domain.

From this figure, one can see a smooth surface with a relatively large plateau. In contrast, for combinations of input values not on the plateau, the relatively steep slope indicates a steep decline in output values for smaller input values. On the plateau one can see, for example, that a rating of 4 in both standards produced the maximum output of teacher performance. In addition, a rating of 4 in Professional Knowledge and a 3 in Assessment Strategies yielded the maximum output as well. It is interesting to note the lack of symmetry in the surface, meaning that a rating of 3 in Professional Knowledge and a rating of 4 in Assessment Strategies did not yield the maximum output but one lower.

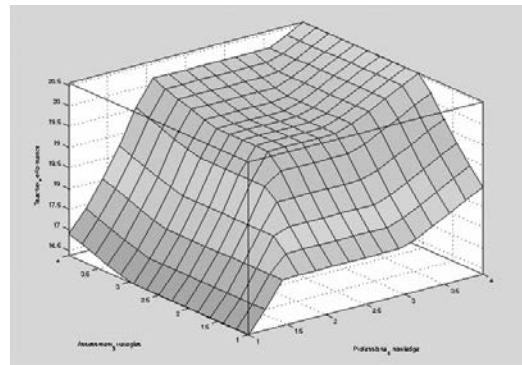


Figure 18. Surface view of Teacher Performance versus Professional Knowledge and Assessment Strategies.

Almost all the surface views for the pairings of two input variables of different domains depicted a surface of this same shape with similar output values. In Figure 19, another example of this type of pairing of input variables is given. As one can see, the surface shows the same shape, but with more symmetrical properties and slightly different output values.

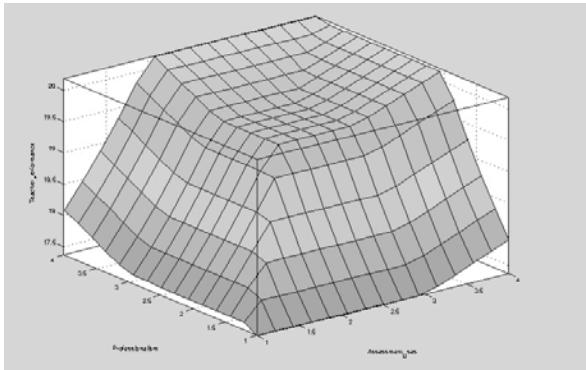


Figure 19. Surface view of Teacher Performance versus Professional Knowledge and Assessment Uses.

In Figure 20, the input variables Professional Knowledge and Instructional Planning, an example of two input variables of the same domain, are plotted. The first observation that can be made is the erratic behavior visible for lower values making apparent ripple-like effects on the visible surface. It is worthy to note that this erratic behavior does not provide inconsistent results for the possible discrete values of the system. The crests of ripples occur at values in between valid input values for the system, which are not possible. The surface maintains consistent outputs for valid input variables, which are the discrete values 1, 2, 3, or 4. What can be observed from the valid inputs is the dramatic difference in output values when both input values are 3 and both input values are 2. This is clearly visible by the relatively steep slope in the center of the

surface. Secondly, one can see the output for the input values 3 and 4 are the same as the output of two 4's. The symmetry of the surface indicates this is independent of assignment of values to the input variable.

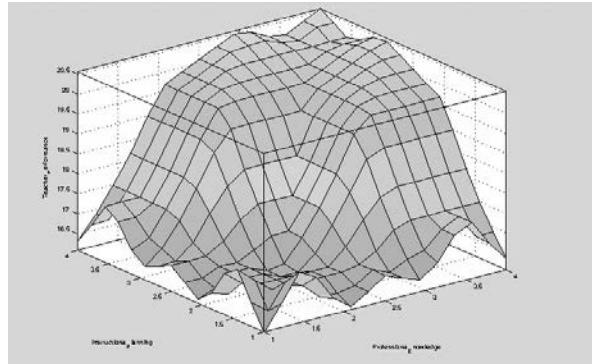


Figure 20. Surface view of Teacher Performance versus Professional Knowledge and Instructional Planning.

As shown in Figure 21, another pairings of standards in the same domain, Assessment Strategies and Assessment Uses, depicts the same shape but a much smoother surface. One may speculate that the difference in smoothness of the surfaces was caused by the weights of the input variables, but further investigation needs to be done to further explain these differences.

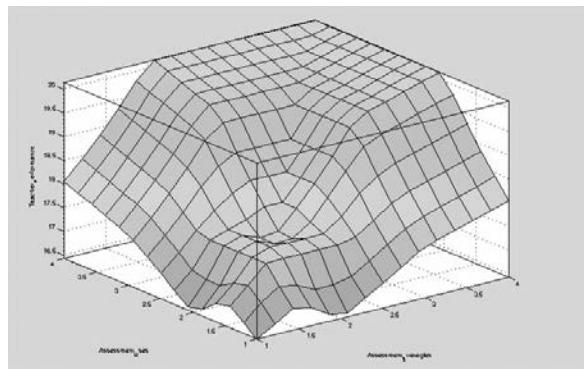


Figure 21. Surface view of Teacher Performance versus Assessment Strategies and Assessment Uses.

The only pairing of standards within the same domain that did not depict this same shape was Instructional Strategies and Differentiation, which is similar to the surface view seen in Figure 22. However, these input variables were also paired with input variables of different domains when the additional rules were written. Figure 23 depicts the general shape of the surface views of the pairing of standards of this type. Erratic behavior is again observed for invalid input values. For an input value of 1.5 for the variable Instructional Strategies, the output values are illogically higher than an input value of 2. Secondly, the asymmetry gives importance to which input variable is assigned a particular value. This effect is the most dramatic when looking at the difference in output values when one input variable's value is assigned a 1, and the other input variable's value is assigned a 4. When the input variable Instructional Strategies is a 1, and the input variable Instructional Planning is a 4, the output value is 15. However, when the input values are reversed, the output value is 18.

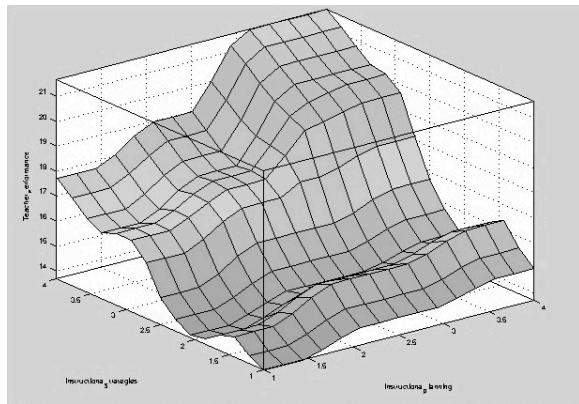


Figure 22. Surface view of Teacher Performance versus Instructional Planning and Instructional Strategies.

Research Question 2

Lack of Variation

Each of 51 teachers who participated in the study received one formative assessment with ratings for each of the 10 standards of the TAPS. The frequency table of the ratings for the individual standards ($n = 510$) whose mean was 3 ($SD = 0.349$) is given in Table 8. As Table 8 indicates, there were few scores given to any one standard that were not a 3, and every teacher in the study received an overall score of 3. Only six individual standards given by the evaluators to five participants were different than a score of 3.

Table 8

Frequency Table of Individual Standard Ratings of Participants ($n = 510$)

Rating	n	%
1	0	0
2	3	0.6
3	504	98.8
4	3	0.6

Estimates of intra-class correlation coefficient (ICC) to measure inter-rater reliability were not calculated because the principals chose not to participate in releasing the data grouped by unnamed evaluators. However, because of such a lack of variability, inter-rater reliability was not a concern.

While the overall rating of each teacher being the same may have been anticipated, the lack of variability within the ratings of the individual standards was surprising. Consequently, five inputs were analyzed in the fuzzy logic expert system as shown in Table 9.

Looking at the five participants in Table 9, Participants 1, 3, and 4 all received one score of 4 or an *Exemplary* rating in one standard. However, Participant 4 received a fuzzy output of 21.4, which was higher than Participants 1 and 3, whose fuzzy output was both 20.6. Participant 2 received a score of 2 in Standard 8, but this did not have an effect on the fuzzy output, which was the same as Participant 6. All other participants not listed in Table 9 received a score of 3 for each rating, and the fuzzy output was 19.6. Lastly, the fuzzy output for Participant 5 was 14.9, which would constitute a different and lower overall rating than the traditional approach. Using the fuzzy logic expert system, Participant 5 would be classified as *Needs Development*; whereas, the traditional approach would classify this participant as *Proficient*.

Table 9

Comparison of the Outputs of the Fuzzy Logic Expert System with the Summative Scores

Participant	Standard										Summative Score	Fuzzy Output
	1	2	3	4	5	6	7	8	9	10		
1	4	3	3	3	3	3	3	3	3	3	21	20.6
2	3	3	3	3	3	3	3	2	3	3	19	19.6
3	3	4	3	3	3	3	3	3	3	3	21	20.6
4	3	3	3	3	3	3	4	3	3	3	21	21.4
5	3	3	2	2	3	3	3	3	3	3	18	14.9
6	3	3	3	3	3	3	3	3	3	3	20	19.6

A comparison of the cut scores for the final ratings for the traditional method and the fuzzy logic expert system are provided in Table 10. As previously stated, the scores for the traditional method are whole numbers. The scores for the fuzzy logic expert system are continuous and follow the conventional rounding method.

Table 10

Comparison of Final Ratings for Traditional and Fuzzy Method.

Final Ratings	TAPS Summative Cut Scores	Fuzzy Logic Outputs
Ineffective	0 – 6	$0 \leq X \leq 6.5$
Needs Development	7 – 16	$6.5 \leq X < 16.5$
Proficient	17 – 26	$16.5 \leq X < 26.5$
Exemplary	27 – 30	$26.5 \leq X < 30$

It was anticipated when acquiring the data from this particular school there would have been a larger sample size than five distinct data points to properly examine how a fuzzy logic expert system can be used to evaluate teacher performance. Consequently, in order to more explicitly answer the second research question, artificial data was incorporated in the study, which is often practiced in this stage of a software development project.

Exemplary Range

The distribution of the summative scores of all possible combinations leading to a rating of exemplary for the ten standards under the traditional method can be seen in Figure 23. The distribution of summative scores is the same as that of a binomial distribution of 10 independent trials with equal probability of success or failure, defined here as an *Exemplary* or *Proficient* rating.

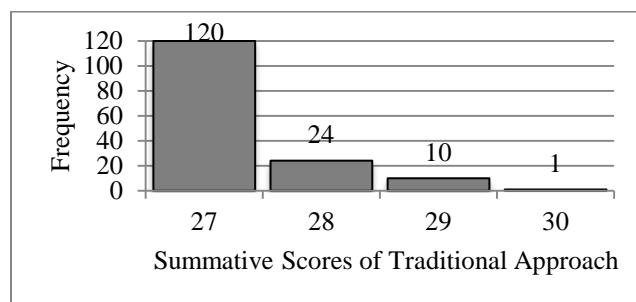


Figure 23. Histograms of the distribution of *Exemplary* ratings of the traditional method.

The distribution of the output values of teacher performance as determined by the fuzzy logic expert system for the same combinations above are shown in Figure 24 with its corresponding frequencies shown in Table 11. The values of the outputs of seventy-two data points (41.1%) ranged from 20-23, no output values ranged from 23-28, and the outputs of one hundred and three data points (58.9%) ranged from 28-30. One can see the distribution of the fuzzy logic expert system is not only different from the traditional method, but it divides all possible combinations into two distinct groups. The group of data points that represented the lower outputs would be classified into the lower category of *Proficient*. Upon closer examination, seven of nine data points in the lowest range of 20 – 21 contained a score of 3 for Standard 3, Instructional Strategies, which was given the most weight in the system.

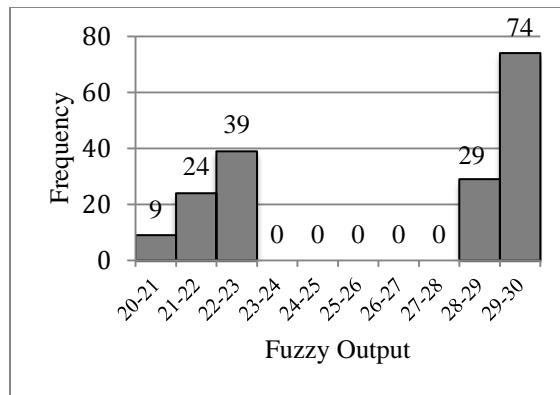


Figure 24. Histogram of the distribution of the fuzzy outputs for all combinations of *Exemplary* ratings of the traditional method.

Table 11

Frequency of the Fuzzy Output for All Combinations of Exemplary Ratings of the Traditional Method.

Characteristic	n	%
Proficient		
20 – 21	9	5.1
21 – 22	24	13.7
22 – 23	39	22.3
Exemplary		
28 – 29	29	16.6
29 – 30	74	42.3

It could be implied from the findings that since the fuzzy logic expert system did not assign a value to any data point that would classify the output in a higher category, the fuzzy logic expert system evaluated more rigorously than the traditional method. However, the results indicated it was also forgiving. There were 74 data points in the highest range, 29 – 30, whereas the traditional method only contained 11 data points in the same interval. The fuzzy output of the data point consisting of all 4's or *Exemplary* ratings was 29.2. The same output was awarded to the set of all data points consisting of

a score of *Proficient* in one standard. Additionally, the set of data points consisting of two *Proficient* ratings when one of these ratings was in Standard 2, Instructional Planning, was given the highest rating, 29.2. Lastly, the results indicated if the data point consisting of two ratings of 3 were for two standards in the same domain, the fuzzy output was rated in the lower category as *Proficient*. The one exception to this case was the data point included in the set previously described when Standard 1, Professional Knowledge, and Standard 2, Instructional Planning, were given a *Proficient* rating.

Upper Range of *Proficient*

During this stage, it was also observed that the fuzzy logic expert system did not assign a fuzzy output that would classify a teacher into a higher category than the one assigned using the traditional summative scores. Consequently, when considering the set of data points of every possible combination of standards for the upper range of *Proficient* (23 – 26) the fuzzy logic expert system did not rate any data point in the higher category, *Exemplary*.

Lower Range of *Proficient*

The distribution of the summative scores of all possible combinations of ratings yielding a summative score in the lower range of *Proficient* for the ten standards under the traditional method can be seen in Figure 25.

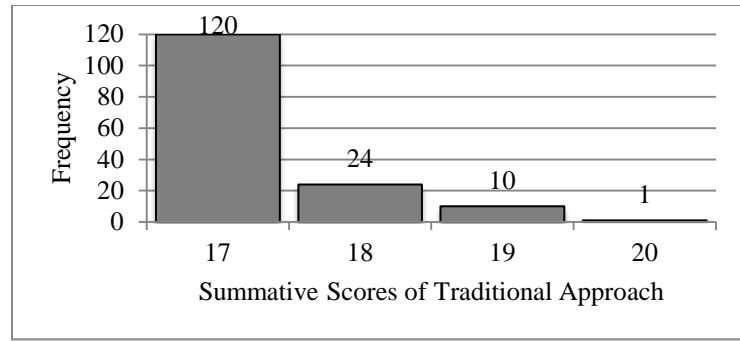


Figure 25. Histogram of the distribution of outputs of the traditional approach.

When examining all data points in the lower range of *Proficient* (17 – 20), the distribution of the values of the outputs from the fuzzy logic expert system is as seen in Figure 26. Table 12 shows the frequency of the class distribution. The output values of sixty-one data points (34.9%) were classified in a lower category, *Needs Development* (7 – 16), and the output values of one hundred fourteen data points (65.1%) were classified as *Proficient* (17 – 26) with the fuzzy logic expert system. The distribution of the fuzzy logic expert system divides all combinations of the data points for this range into three groups, although the distinction or range of values where no output values lie was not as large as those for the set of data points previously described for the *Exemplary* range. The three groups did not mark the boundaries for classification, however. The middle group of output values is classified into both *Needs Development* and *Proficient*. The fifteen data points (8.6%) whose output values ranged from 16 – 17 were classified as *Needs Development*. The forty-nine data points (28.0%) whose output values ranged from 17 – 18 were classified as *Proficient*.

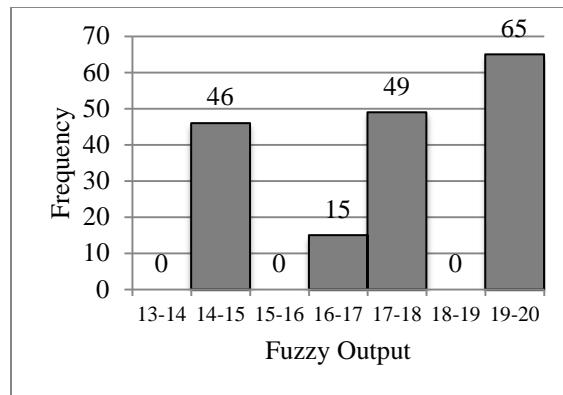


Figure 26. Histogram of the distribution of the outputs of the fuzzy logic method.

Table 12

Frequency of Class Distribution of the Fuzzy Output.

Characteristic	n	%
Needs Development		
14 – 15	46	26.3
15 – 16	0	0
16 – 17	15	8.6
Proficient		
17 – 18	49	28.0
18 – 19	0	0
19 – 20	65	37.1

Closer examination of the 46 data points whose output value was from 14 – 15 revealed that in each of these cases a score of 2 was received for Standard 3, Instructional Strategies. The data point consisting of all 3's but one 2 in Standard 3 was also in this interval. While this standard was assigned the most weight, this data point was also calculated when all rules were weighted equally, and the output was 15.4, which would still classify as *Needs Development*.

Conclusion

While there was little variability in the data obtained from the high school in the case study, the proof of concept responded to both the artificial and real world data, which indicated that the fuzzy logic expert system was capable of producing valid outputs to evaluate teacher performance. The artificial data from the prototype stage indicated a different class distribution in the fuzzy logic expert system. Distinct groups were identified for those teachers whose performance scores were at the border of two different ratings. The fuzzy logic expert system did not award data points to a class higher than those determined by the traditional approach, but the highest fuzzy output possible was given to a much larger number of teachers than with the traditional approach.

CHAPTER SIX

CONCLUSIONS AND IMPLICATIONS

This chapter will provide a brief summary of the study, discuss implications of the results relating the findings to prior research, and provide recommendations and suggestions for future studies. The purpose of this study was to examine to what degree a fuzzy logic expert system could elicit and quantify teacher performance using state teacher evaluation methods. To accomplish that goal, the basis of the rules for the fuzzy logic expert system were developed from the decision-making matrices provided by the state, and the weights for the rules were assigned based on a survey given to the four administrators at an unidentified high school. Additional rules were written based on a follow-up interview with the administrators. Ratings on the ten Teacher Assessment Performance Standards were collected from participating certified teachers and evaluated by the traditional method and the fuzzy logic method. Further comparisons were made by analyzing the distributions of the set of all possible ratings for summative scores lying at the boundaries of two different classifications of teacher performance.

The lack of variation in the ratings given by the evaluators at the school used in the study did not allow for a complete picture comparing the results of the summative scores of the fuzzy logic expert system to the current state evaluation method. Thus, the

fuzzy logic expert system was examined further through an analysis of distributions of all possible ratings at the boundary of the *Proficient* and *Exemplary* categorical ratings of teacher performance. The results indicated that the fuzzy logic expert system was able to make a distinction between results lying at these boundaries of performance categories.

Implications of a Fuzzy Logic Expert System

A Fuzzy Logic Expert System for Quantifying Teacher Performance

The results of this study are similar to those found by Bhosale and Kamath (2013), Chaudhari et al. (2012), and Trsetenjak and Donko (2013), which claimed that a fuzzy logic expert system provided a more advanced and sensitive approach in the teacher performance evaluation process that could ultimately result in a more realistic evaluation. In contrast to prior studies, the findings of this study provided examples to support these generalized claims. The findings in this study indicated that the fuzzy logic expert system did not assign a value to any data point that would classify the output in a higher category.

The surface views of teacher performance showed distinct areas of performance with smooth transitions from one category to the next similar to the study by Jamsandekar and Mudholkar (2013). These results indicated that this smooth distribution helped to identify teachers lying at the border of two class distributions. Furthermore, from the degree of the slope in the surfaces related to the written rules, the rate at which the outputs were classified from one category to the next could be determined.

For example, in the case of two standards paired from different domains, one observed a smooth surface with a relatively large plateau. For combinations of input

values on the plateau, the output values remain high. For combinations of input values not on the plateau, the steepness in the slope indicates the rate of decrease in output values for smaller input values. In the case of two standards paired from the same domain, the area of the plateau is smaller, meaning the output values begin to decrease as the values of the input variables decrease more rapidly than the previous case. In the case of two standards from different domains that were paired with the additional rules written, the individual weight of the rules affects the slope of the surface and thus affecting the symmetry of the surface.

Distinguishing Between Performance Categories in the Fuzzy Logic Expert System

Although other studies in the Eastern World (Bhosale & Kamath, 2013; Djam & Mishra, 2013; Jamsandekar & Mudholkar, 2013; Atta-ur-Rahman, 2013) have used fuzzy logic to build an evaluation system for teacher performance, this study is the first done in the United States with present teacher evaluation methods. While these results cannot be generalized, the fuzzy logic expert system demonstrated the ability to separate values of teacher performance at the boundary of a category of performance into two distinct groups. For example, for the set of all data points that classify as *Exemplary* under the traditional system, the fuzzy logic expert system divided the set into two distinct groups. Forty-one percent of data points ranged from 20 – 23, 59% of data points from 28 – 30, and no data points were in the middle of these ranges. This ability shows the potential for a fuzzy logic expert system to provide a more accurate continuum of teacher performance.

Secondly, the results indicated that the fuzzy logic expert system was able to distinguish between teachers, who may exhibit exemplary qualities in those standards

deemed most important by administrators, and those teachers, who may exhibit exemplary qualities in those less important. For example, the output of the fuzzy logic expert system was lower for input values of a *Proficient* rating for Standard 3, Instructional Strategies, rated as most critical to teacher performance, than input values of a *Proficient* rating for Standard 10, Communication, which was rated less critical.

Distinguishing Between Performance Categories in the Traditional Method

Although it could be implied that the fuzzy logic expert system evaluated more rigorously than the traditional method, it also demonstrated the ability to be forgiving in many cases. For example, the output value of teacher performance was the same for input values of a teacher who may receive all *Proficient* ratings and another who may receive nine *Proficient* ratings and one *Needs Development*. This quality of the system could have many implications for school administration and leaders. Perhaps this could encourage evaluators to differentiate among the ratings of individual standards to provide meaningful feedback to support growth in a teacher's performance if they were aware that a summative score would not be affected.

While the overall rating of each teacher being the same may have been somewhat anticipated, the lack of variability within the ratings of individual standards was not. The ratings at this school appear to be influenced by compression and benchmarking, restricting the evaluation score's range of variation as seen in the review of literature under previous evaluation systems (Barrett et al., 2014; Batten, 2013; Hill et al., 2012; Ho & Kane, 2013; Jacob & Lefgren, 2008). Such benchmarking occurs when there is a

discrete rating that is designated as a cutoff for acceptable performance (Barrett et al., 2014). The data indicated the evaluators used the *Proficient* rating as a benchmark, where every teacher was awarded the second to highest rating. This benchmarking effect may have been caused by the following factors. The Georgia Department of Education states that the *Proficient* rating is the target and emphasizes the *Exemplary* rating should not be easy to attain (Georgia Department of Education, 2013b). In addition, a Professional Development Plan (PDP) will be required of a teacher if the overall Teacher Effectiveness Measure is rated *Needs Development*, which may have discouraged evaluators from assigning this rating, even if the *Proficient* rating target was not observed.

The results were surprising because such evidence of a benchmarking effect for the newly adopted teacher performance evaluation system was not anticipated at the time of the study. However, the findings at this particular school are consistent with recently released empirical evidence found across Georgia and other states that have adopted a new teacher performance evaluation system (Anderson, 2013; Daly, 2014). For the 2012 – 2013 assessment data of school districts participating in the pilot study, 96.9 percent of participating teachers scored *Proficient* or *Exemplary* (Georgia Department of Education, 2014). DeKalb County, in the 2013 – 2014 pilot study, had 83 exemplary teachers among its 6,500 teachers (1%), but the number of teachers rated as *Proficient* were not reported (Downey, 2014). Similarly, the Indiana Department of Education reported 97.3 percent of teachers were rated as “effective” or “highly effective.” In Florida, 97 percent of teachers and in Michigan, 98 percent of teachers belonged in these two categories, and in Tennessee, 98 percent of teachers were “at expectations” (Anderson, 2013). However,

the released data previously mentioned does not disclose the ratings of the individual standards to compare the findings at the school in this study with those in these reports.

While teachers may no longer achieve the highest score, there is evidence that they are still receiving the same score. As Daly (2014), President of the New Teacher Project, frankly stated, “You can’t fix evaluations if observers don’t rate accurately” (p. 1). After spending millions of dollars and thousands of hours of training to develop new evaluation systems to better distinguish levels of teacher performance, the observation ratings still bear a striking resemblance to the Widget Effect, where almost every teacher receives the same rating (New Teacher Project, 2013). If every teacher receives the same rating for each standard, no distinction can be made. Because this appears to remain the case with new teacher evaluation systems, the fuzzy logic expert system had a small sample size to show its potential in the field of teacher performance. Within this small sample, the fuzzy logic expert system rated a participant as a different category of performance than the traditional approach.

Addressing and correcting the lack of variability takes review of rating distributions by the evaluators and district wide discussion of the correlation between ratings and student outcomes (Daly, 2014; New Teacher Project, 2013). Most school districts hold school leaders responsible only for basic compliance tasks when it comes to observations (conducting the minimum number of classroom visits or submitting observation forms on time) (New Teacher Project, 2013). Perhaps if principals’ evaluations were linked to student performance, this would encourage principals to evaluate more rigorously and accurately (Goldhaber, 2007). McKay (2013) suggests in order for school leaders to help teachers overcome their fear of constructive feedback,

leaders must create a culture of learning in which a teacher does not have to be bad to be better.

Suggestions for Further Research

Reasons for Lack of Variability in New Teacher Performance Evaluation Systems

The results from this study, indicating that the new teacher evaluation system is failing to distinguish levels of performance among teachers, has opened many new avenues to explore in future studies. The lack of variability can be speculated to have been the result of several factors or constraints. Researchers have reported when evaluators have too much to look for in a short amount of time, they rate most of the indicators similarly, and the observations begin to resemble the superficial checklists of the past (Donaldson, 2009; Ho & Kane, 2013; Kane & Staiger, 2012). Other constraints, which could have contributed to the observed ratings, are school culture resistant to less than perfect evaluation ratings or few to no incentives for administrators to evaluate accurately (Donaldson, 2009). Similarly, the compression of scores could have been the result of the principals' sense of fairness or a desire to avoid confrontation during feedback as seen in other studies (Edenfield, 2014; Salleh, Amin, Muda, Sofian, & Halim, 2013). Future research could further examine these reasons, among others, why there continues to be a lack of variation in principal evaluations of the new teacher evaluation systems.

This pattern of more lenient and less variance in ratings has been reported to be even greater when the principal evaluations are used for high-stake decisions such as

teacher compensation (Dee & Wyckoff, 2015; Miller, 2009). The prevalence of null findings from this study, among others, raises considerable doubt and resistance about the use of newly developed teacher evaluation systems to make high-stake decisions related to issues such as compensation or dismissal (Dee & Wyckoff, 2015; Edenfield, 2014; Weiss & Long, 2013; Zubrzycki, 2012). Such could be the case for principals in Georgia where new policy is using these evaluation scores for high stakes decisions (Rickman & Olivarez, 2014). In the fall of 2014, Georgia began compensation redesign, which stated their commitment to pay exemplary teachers the most money (Georgia Department of Education, 2014b). In addition, teachers who receive two Teacher Effectiveness Measures of *Ineffective* or *Needs Development* within a five-year period will be unable to renew their teaching certificate (Georgia Department of Education, 2013b). Perhaps it is these state policies themselves that consequently inhibit evaluators' efforts to differentiate among teachers. Future research could investigate the evaluators' perceptions of the new evaluation systems or examine the relationship between high-stakes decisions and lack of variation in principal evaluations of teacher performance.

Incorporate Fuzzy Neural Networking Techniques

A larger data collection, which incorporates a large number of schools, would perhaps not only provide a greater range of variation in the teacher performance ratings but also allow researchers to explore other fuzzy logic techniques such as clustering and other fuzzy artificial neural network techniques to train the model and find classes of patterns to refine the rules of the system (Nasira, Kumar, & Kiruba, 2008; Neogi et al., 2011). Future studies could incorporate an analytical hierarchy process, another artificial

neural networking fuzzy technique, which systematically and efficiently increases the number of written rules number since the number of rules naturally increases exponentially (Lee et al., 2003).

Confidential Principal Evaluations and Values on a Continuum

In order to design a fuzzy logic expert system that mimicked the present ratings of the Georgia TKES, the range of input values was restricted to the values of 1 – 4. Each of these values represented the four categories of teacher performance. For example, a value of 3 restricted a teacher's performance in one standard to entirely belong to the rating *Proficient* or not at all. Future studies could expand the range of input values for the rating of each standard to values of 1 – 12, for example. This would allow further flexibility for an evaluator to rate a teacher as somewhat *Exemplary* or very *Proficient*. Then the system would be able to employ an important advantage of implementing a fuzzy logic expert system, which allows the membership of the rating to belong to performance categories as a matter of degree.

Secondly, in contrast to using the data used for the teacher performance measure of the new teacher performance evaluation systems, future studies could incorporate ratings of the state evaluation rubrics where the evaluations of principals were completely confidential and not used for any type of performance measure, as in the cases of the studies by Harris and Sass (2009) and Jacob and Lefgren (2008). Perhaps, there would be more variation in the ratings of the standards by the evaluators if teachers did not see the ratings. This would allow a larger sample size of participants with different individual ratings to evaluate with a fuzzy logic expert system.

A Fuzzy Logic Expert System Using a Value-added model

In the pilot year 2014 – 2015 for the full implementation of Teacher Keys

Effectiveness System (TKES) for every district in Georgia, student achievement scores

will not be used or available to calculate the overall Teacher Effectiveness Measure.

However, as student growth data becomes available in Georgia and other states, it would also be worth further research to incorporate student growth into the system. One way to incorporate a value-added model into a fuzzy logic expert system design could be to write rules based on both the results of student growth and the observation scores given by evaluators. This would be valuable to study because most new teacher evaluation systems will incorporate both observational measures of instruction and measures of student achievement gains (Hull, 2013; Kane et al., 2014).

Conclusion

A majority of states have implemented new reforms in teacher evaluation strategies in order to diagnose and improve teacher performance, but unfortunately, have not incorporated a method capable of accurately representing qualitative factors and the relationships between them. In order to provide decision makers with a more accurate measure of teacher performance evaluation, there is a need for the application of techniques such as fuzzy logic to model and measure teacher performance. Although other studies in the Eastern World have used fuzzy logic to build a prototype for evaluating teacher performance, this study was the first to be done in the United States with current teacher evaluation methods. The results of this study can be used to help researchers better understand how a fuzzy logic expert system can quantify teacher performance and identify and distinguish between levels of performance. It is hoped that

this study will further understanding and help bridge the gap that currently exists between fuzzy logic and education.

References

Ahmad, H. & Asri, N. (2013). In pursuing better academic result in university: A case of fuzzy logic analysis. In proceeding of the 2013 International Conference on Education and Modern Educational Technologies, 98-102.

Ajiboye, A.B. & Weir, R.F. (2005). A heuristic fuzzy logic approach to EMG pattern recognition for multifunctional prosthesis control. *Neural Systems and Rehabilitation Engineering* 13(3), 280-291. doi:10.1109/TNSRE.2005.847357.

Al Ganideh, S.F., El Refae, G.A., & Aljanaideh, M.M. (2011). Can fuzzy logic predict consumer ethnocentric tendencies? An empirical analysis in Jordan. Paper presented at the 2011 Annual Meeting of the North American Fuzzy Information Processing Society. El Paso, Texas. doi:10.1109/NAFIPS.2011.5752001.

Almy, S. (2011). Fair to everyone: Building the balanced teacher evaluations that educators and students deserve. Educationtrust. Retrieved from http://www.edtrust.org/sites/edtrust.org/files/Fair_To_Everyone_0.pdf.

Amin, H. & Khan, A. (2009). Acquiring knowledge for evaluation of teachers' performance in higher education using a questionnaire. *International Journal of Computer Science and Information Security*, 2(1), 1-7.

Ammar, S., Bifulco, R., Duncombe, W., & Wright, R. (2000). Identifying low performance public schools. *Studies in Educational Evaluation*, 26(3), 259-287.

Anderson, J. (2013). Curious grade for teachers: Nearly all pass. *The New York Times*, March 30, 2013.

Atta-ur-Rahman, M. (2013). Teacher assessment and profiling using fuzzy rule based system and Apriori algorithm. *International Journal of Computer Applications*, 65(5), 22-28.

Bai, S. & Chen, S. (2008a). Evaluating students' learning achievement using fuzzy membership functions and fuzzy rules. *Expert Systems with Applications*, 34(2008), 399-410.

Bai, S. & Chen, S. (2008b). Automatically constructing grade membership functions of fuzzy rules for students' evaluation. *Expert Systems with Applications*, 35(2008), 1408-1414.

Barnett, J.H., Rinthapol, N., & Hudgens, T. (2014). TAP research summary: Examining the evidence and impact of TAP: The system for teacher and student advancement. *National Institute for Excellence in Teaching*. Retrieved from <http://www.niet.org/assets/PDFs/tap-research-summary.pdf>.

Barrett, N., Crittenden-Fuller, S., & Guthrie, J. (2014). *Subjective ratings of teachers: Implications for strategic and high-stakes decisions*. Paper presented at 39th Annual Conference of the Association for Education Finance and Policy, San Antonio, Texas.

Batten, D. (2013). *Teacher performance evaluations and value-added scores: Evidence from North Carolina Public Schools*. (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill, NC.

Bhosale, G.A. & Kamath, R.S. (2013). Fuzzy inference system for teaching staff performance appraisal. *International Journal of Computer and Information Technology*, 2(3), 381-385.

Bill & Melinda Gates Foundation. (n.d.). *Foundation commits \$335 million to promote effective teaching and raise student achievement*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from [www.gatesfoundation.org/Media-Center/Press-Releases/2009/11/Foundation-Commits-\\$335-to-Promote-Effective-teaching-and-Raise-Student-Achievement](http://www.gatesfoundation.org/Media-Center/Press-Releases/2009/11/Foundation-Commits-$335-to-Promote-Effective-teaching-and-Raise-Student-Achievement).

Biswas, R. (1995). An application of fuzzy sets in students' evaluation. *Fuzzy Sets and Systems*, 74(1995), 187-194.

Bjelica, M. & Rankovic, D. (2010). The use of fuzzy theory in grading of students in math. *Turkish Online Journal of Distance Education*, 11(1), 13-19.

Bouchon-Meunier, B., Dotoli, M., & Maione, M (1996). On the choice of membership functions in a Mamdani-type fuzzy controller. Paper presented at the First Online Workshop on Soft Computing, Nagoya, Japan, 1996.

Cantrell, S. & Kane, T.J. (2013). Measures of Effecting Teacher Project (2013). *Ensuring fair and reliable measures of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

Carlo, M. (2011, July 19). *Teacher evaluations: Don't begin assembly until you have all the parts*. Retrieved from <http://shankerblogger.org/?p=3165>.

Chang, D.F. & Sun, C.M. (1993). Fuzzy assessment of learning performance of junior high school students. In Proceedings of the 1993 first national symposium on fuzzy theory and applications, Hsinchu, Taiwan, Republic of China, 1-10.

Charles, V., Kumar, M., & Suggu, S. (2013). Adapting fuzzy linguistic SERVQUAL model: A comparative analysis of bank services in Malaysia. Working paper CENTRUM Catolica-Pontificaia Universida Catolica del Peru.

Chaudhari, O.K., Khot, P.G., & Deshmukh, K.C. (2012). Soft computing model for academic performance of teachers using fuzzy logic. *British Journal of Applied Science & Technology*, 2(2), 213-226. Retrieved from www.sciedomain.org.

Chen, S. & Lee, C. (1999). New methods for students' evaluation using fuzzy sets. *Fuzzy Sets and Systems*, 104(1999), 209-218.

Chen, S. & Wang H. (2009). Evaluating students' answerscripts based on interval-valued fuzzy grade sheets. *Expert Systems with Applications*, 36 (2009), 9839-9846.

Cheng, Y. & Chung, M. (2014). A new measurement method on correlation coefficient for attribute fuzzy interval data and its applications. *International Journal of Intelligent Technologies & Applied Statistics*, 7(1), 27-36.
doi:10.6148/IJITAS.2014.0701.03.

Cheng, C., Wang, J., Tsai, M., & Huang, K. (2004). Appraisal support system for high school teachers based on fuzzy linguistic integrating operation. *Journal of Human Resource Management*, 4(2004), 73-89.

Cheng, Y. & Yang, C. (2013). The application of fuzzy correlation coefficient with fuzzy interval data. *International Journal of Innovative Management, Information & Production*, 4(1), 65-71.

Chiang, T.T. & Lin, C.M. (1994). Application of fuzzy theory to teaching assessment. In Proceedings of the 1994 second national conference on fuzzy theory and applications, Taipei, Taiwan, Republic of China, 92–97.

Chukwubikem, E. (2012). Developing better teacher evaluation. *International Journal of Social Sciences & Education*, 2(4), 554-566.

Cole, J.R. & Persichitte, K.A. (2000). Fuzzy cognitive mapping: Applications in education. *International Journal of Intelligent Systems*, 15(1), 1-25.

Connecticut Department of Education. (2014). *SEED: Connecticut's system for educator evaluation and development*. Retrieved from http://www.connecticutseed.org/wpcontent/uploads/2012/10/SEED_Handbook.pdf.

Corcoran, S.P. (2010). Can teachers be evaluated by their students' test scores? Should they be? Executive summary. Education Policy for Action Series. *Annenberg Institute for School Reform at Brown University*.

Daley, G. & Kim, L. (2010). A teacher evaluation system that works. *National Institute for Excellence in Teaching*. Working paper.

Daly, T. (2014). *4 Things we've learned since the Widget Effect*. New Teacher Project. Retrieved from <http://tntp.org/blog/post/4-big-things-weve-learned-about-teacher-evaluation-since-the-widget-effect>.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dee, T. & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2), 267–297.

Djam, X.Y. & Mishra, A.K. (2013). Fuzzy cognitive map based approach for teachers' performance evaluation. *Pacific Journal of Science and Technology*, 14(2), 176-181.

Donaldson, M. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/issues/2009/06/pdf/teacher_evaluation.pdf.

Donaldson, M. (2012). *Teachers' perspectives on evaluation reform*. Center for American Progress. Retrieved from <https://www.americanprogress.org/issues/education/report/2012/12/13/47689/teachers-perspectives-on-evaluation-reform/>.

Downey, M. (2014, April 3) State approves education evaluations that weigh student performance. *Atlanta Journal Constitution*. Retrieved from <http://www.ajc.com/weblogs/get-schooled/2014/apr/03/state-approves-education-evaluations-weigh-student/>.

Duffett, A., Farkas, S., Rotherham, A.J., & Silva, E. (2008). *Waiting to be won over: Teachers speak on the profession, unions, and reform*. Retrieved from <http://all4ed.org/reports-factsheets/waiting-to-be-won-over-teachers-speak-on-the-profession-unions-and-reform/>.

Durkin, J. (1990). Research review: Application of expert systems in the sciences. *The Ohio Journal of Science*, 90(5), 171-179.

Echauz, J.R. & Vachtsevanos, G.J. (1995). Fuzzy grading system. *IEEE Transactions on Education*, 38(2), 158- 165.

Edenfield, J. (2014). *Teachers' perceptions of merit pay in Georgia*. Electronic Theses & Dissertations. Paper 1058. (Doctoral dissertation).

Edwards, W. & Newman, J.R. (1982). *Multiattribute evaluation*. Beverly Hills, CA: Sage Publications.

Exstrom, M. (2013). *Do teachers make the grade?* Retrieved from <http://www.ncsl.org/research/education/do-teachers-make-the-grade.aspx>.

Fagan, L.M. (1978). Ventilator manager: A program to provide on-line consultative advice in the intensive care unit. Retrieved from <https://saltworks.stanford.edu/catalog/druid:rd792fk1120>.

Fourali, C. (1997). Using fuzzy logic in educational measurement: The case of portfolio assessment. *Evaluation and Research in Education*, 11(3), 129-148.

Gallagher, H.A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, 79(4), 79–107.

Ganideh S.F. & Aljanaideh, M. (2013). Using fuzzy logic to analyze marketing data: The impact of socio-psychological variables on the national identity of Jordanians. *Transnational Corporations Review*, 5(2), 104-114.

Ganideh, S., Refae, G., & Aljanaideh, M. (2011). Can fuzzy logic predict consumer ethnocentric tendencies? An empirical analysis in Jordan. *Journal of Physical Science and Application*, 1(2011), 100-106.

Garibaldi, J. & John, R. (2003). Choosing membership functions of linguistic terms.

Paper Presented at the 2003 IEEE International Conference on Fuzzy Systems. St. Louis, USA, 578–583.

Georgia Department of Education. (2011). *CLASS KeysSM: Classroom Analysis of State Standards: The Georgia Teacher Evaluation System*. Atlanta, GA. Retrieved from <http://www.gadoe.org/School-Improvement/Teacher-and-LeaderEffectiveness/Documents/CLASSLeader%20Keys/CK%20Process%20Guide%203-23-2011.pdf>.

Georgia Department of Education. (2013a). *Teacher Keys and Leader Keys Effectiveness System: 2012 pilot evaluation report*. Retrieved from <http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/Pilot%20Report%2012-13-2012%20FINAL%20Clean.pdf>.

Georgia Department of Education. (2013b). *Teacher Keys Effectiveness System*. Retrieved from <https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Pages/Teacher-Keys-Effectiveness-System.aspx>.

Georgia Department of Education. (2014a). *Overview/executive summary of the 2012-2013 TKES and LKES evaluation report*. Retrieved from http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/FINAL%20Year%203%20Report%20_2-21-2014_FORMATTED%202-23-2014.pdf.

Georgia Professional Standards Commission (2014b). *Race to the Top: Georgia's vision for educational excellence*. Retrieved from <http://www.gadoe.org/Race-to-the-Top/Documents/Race%20to%20the%20Top%20Four%20Year%20Report%20by%20GPEE.pdf>.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved from <http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf>.

Goldhaber, D. (2007). Principal compensation: More research needed on a promising reform. Center for American Progress. Retrieved from https://cdn.americanprogress.org/wpcontent/uploads/issues/2007/12/pdf/principal_pay.pdf.

Goldhaber, D. (2009). Exploring the use of incentives to influence the quality and distribution of teachers. In S. Sclafani (Ed.), *Evaluating and Rewarding the Quality of Teachers: International Practices*. Organization of Economic Development and Cooperation, OECD.

Goldhaber, D. & Loeb, S. (2013). *What do we know about the tradeoffs associated with teacher misclassification in high stakes personnel decisions?* The Carnegie Knowledge Network. Retrieved from <http://cepa.stanford.edu/content/what-do-we-know-about-tradeoffs-associated-teacher-misclassification-high-stakes-personnel-decisions>.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington, DC: Center for Educator Compensation Reform. Retrieved from http://cecr.ed.gov/pdfs/Inter_Rater.pdf.

Gupta, L. & Dwahan, A. (2012). Diagnosis, modeling and prognosis of learning system using fuzzy logic and intelligent decision vectors. *International Journal of Computer Applications*, 37(6), 25-29.

Hallinger, P., Heck, R., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, 2014(26), 5-28.

Hansen, M., Lemke, M., & Sorensen, N. (2013). *Combining multiple performance measures: Do common approaches undermine districts' personnel evaluation systems?* Washington, DC: American Institutes for Research. Retrieved from http://www.air.org/files/Combining_Multiple_Performance_Measures.pdf.

Harris, D.N. & Sass, T.R. (2009). What makes for a good teacher and who can tell? Working Paper 30. National Center for Analysis of Longitudinal Data in Education Research.

Hill, H.C., Charalambous, C.Y., & Kraft, M.A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41-56. doi: 10.3102/0013189X12437203.

Hill, H.C. & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, 83(2), 371-384.

Ho, A.D. & Kane, T.J. (2013). *The reliability of classroom observations by school personnel*. Research Paper. MET Project. Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Reliability_of_Classroom_Observations_Research_Paper.pdf.

Hong, T. & Lee, C. (1996). Induction of fuzzy rules and membership functions from training examples. *Fuzzy Sets and Systems*, 84(1), 33-47.

House, M. (2012). Fuzzy logic-based democracy index. Paper presented at the 2012 ACMSE conference. Tuscaloosa, AL.

Huapaya, C.R. (2012). *Proposal of fuzzy logic-based students' learning assessment model*. Retrieved from <http://sedici.unlp.edu.ar/bitstream/handle/10915/23652/4756-Proposal%20of%20Fuzzy%20Logic%20based%20Students%20Learning%20Assessment%20Model.pdf?sequence=1>.

Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance*. Center for Public Education. Retrieved from <http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf>.

Ingoley, S. & Bakal, J.W. (2012). Use of fuzzy logic in evaluating students' learning achievement. *International Journal on Advanced Computer Engineering and Communication Technology*, 1(2), 47-56.

Jacob, B.A. & Lefgren, L. (2005). Principals as agents: Subjective performance measurement in education. Working Paper 11463. National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w11463>.

Jacob, B.A. & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1), 101–136.

Jacob, B.A., Lefgren, L., & Sims, D.P. (2008). The persistence of teacher-induced learning. *Journal of Human Resources*, 45(4), 915-943.

Jamsandekar, S. & Mudholkar, R.R. (2013). Performance evaluation by fuzzy inference technique. *International Journal of Soft Computing and Engineering*, 3(2), 158-164.

Joe, J., Tocci, C., Holtzman, S., & Williams, J. (2013). *Foundations of observation. Measures of Effective Teaching (MET) Project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from www.metproject.org/.../MET-ETS_Foundations_of_Observation.pdf.

Kane, T.J. & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Measures of Effective Teaching (MET) Project. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

Kane, T.J., Kerri, K.A., & Pianta, R.C. (2014). *Designing Teacher Evaluation Systems*.

New Jersey: John Wiley & Sons, 1-674.

Kane, T.J., McCaffrey, D.F., Miller, T., & Staiger, D.O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. MET Project. Bill & Melinda Gates Foundation. Retrieved from http://www.hec.ca/iea/seminaires/140401_staiger_douglas.pdf.

Kane, T.J., Rockoff, J.E. & Staiger, D.O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, 27(6), 615-631.

Kane, T.J. & Staiger, D.O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Measures of Effective Teaching (MET) Project. (2012). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2011). Evaluating teacher effectiveness. *Education Next*, 11(3), 54-60.

Kannemeyer, L. (2005). Reference framework for describing and assessing students' understanding in first year calculus. *International Journal of Mathematical Education in Science & Technology*, 36(2/3), 271-287.

Kao, Y., Lin, Y., & Chu, C. (2012). A multi-factor fuzzy inference and concept map approach for developing diagnostic and adaptive remedial learning systems. *Procedia - Social and Behavioral Sciences*, 64 12th International Educational Technology Conference – IETC, 2012, 65-74. doi:10.1016/j.sbspro.2012.11.009.

Khan, A.R., Amin, H.U., & Rehman, Z.U. (2011). Application of expert system with fuzzy logic in teachers' performance evaluation. *International Journal of Advanced Computer Science and Application*, 2(2), 51-57.

Khan, S. & Quaddus, M. (2004). Group decision support using fuzzy cognitive maps for causal reasoning. *Group Decision and Negotiation*, 13, 463-480.

Kimball, S.M. & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, 45(1), 34-70.

Kosko, B. (1986). Fuzzy cognitive maps. *International Journal of Man-Machine Studies*, 24, 65-75.

Kosko, B. (1993). *Fuzzy thinking: The new science of fuzzy logic*. New York: Hyperion.

Kosko, B. & Isaka, A. (1993). Fuzzy logic. *Scientific American*, 1993, 76-81.

Kumar, O.S. (2013). A fuzzy based comprehensive study of factors affecting teacher's performance in higher technical education. *International Journal of Modern Education & Computer Science*, 5(3), 26-34. doi:10.5815/ijmecs.2013.03.04.

Law, C. (1996). Using fuzzy numbers in educational grading system. *Fuzzy Sets and Systems*, 83(1996), 311-323.

Lee, M., Chung, H., & Yu, F. (2003). Modeling of hierarchical fuzzy systems. *Fuzzy Sets & Systems*, 138(2), 343-355. doi:10.1016/S0165-0114(02)00517-1.

Lemmon, H. (1986). COMAX: An expert system for cotton crop management. *Science* 233, 29-33.

Ma, J. & Zhou, D. (2000). Fuzzy set approach to the assessment of student-centered learning. *IEEE Transactions on Education*, 43(2), 237–241.

Mandal, S.N., Choudhury, J., De, D., & Chaudhuri, S.B. (2008). Roll of membership functions in fuzzy logic for prediction of shoot length of mustard plant based on residual analysis. *World Academy of Science, Engineering and Technology*, 38, 378-384.

Maslow, V.J. & Kelley, C. J. (2012). Does evaluation advance teaching practice? The effects of performance evaluation on teaching quality and system change in large diverse high schools. *Journal of School Leadership*, 22(3), 600-632.

Massey (2012). *Fuzzy logic* [PowerPoint slides]. Retrieved from <http://www.massey.ac.nz/~nhreyes/MASSEY/159741/Lectures/Lec2012-3-159741-FuzzyLogic-v.2.pdf>.

Mathworks, Inc. (2013). *Fuzzy logic toolbox user's guide*. Retrieved from www.mathworks.com.

McKay, C. (2013). No more Lake Wobegon. *You don't have to be bad to get better*. Corwin: Thousand Oaks, California: Corwin Press, 19-38.

Measures of Effective Teaching (MET) Project. (2010). *Working with teachers to develop fair and reliable measures of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from www.metproject.org/downloads/met-framing-paper.pdf.

Measures of Effective Teaching (MET) Project. (2010). *Validation engine for observational protocols*. Seattle, WA: Bill & Melinda Gates Foundation.

Retrieved from
http://www.metproject.org/downloads/Validation_Engine_concept_paper_092410.pdf.

Measures of Effective Teaching (MET) Project. (2013). *Feedback for better teaching: Nine principles for using measures of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from
http://metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles %20Paper.pdf.

Medley, D.M. & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, 80(4), 242–247.

Mendel, J.M. (1995). Fuzzy logic systems for engineering: A tutorial. *IEEE Journal*, 83, 345–377.

Mihaly, K., McCaffrey, D.F., Staiger, D.O., & Lockwood, J.R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from
http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, 79(4), 33–53.

Milanowski, A. (2011). Strategic measures of teacher performance. *Phi Delta Kappan*, 92(7), 19–25.

Milanowski, A.T., Kimball, S.M., & White, B. (2004). The relationship between standards-based teacher evaluation scores and student achievement. Consortium for Policy Research in Education, University of Wisconsin-Madison. Retrieved from http://cpre.wceruw.org/papers/3site_long_TE_SA_AERA04TE.pdf.

Milanowski, A.T., Prince, C.D., & Koppich, J. (2007). *Observations of teachers' classroom performance*. Center for Educator Compensation Reform. Retrieved from <https://www.wested.org/wp-content/uploads/classroom-practice.pdf>.

Miller, T. (2009). *Investigation elementary teachers' perceptions about and experiences with Ontario's teacher performance appraisal system*. (Doctoral dissertation). Retrieved from
https://tspace.library.utoronto.ca/bitstream/1807/19155/3/Miller_Thomas_J_2009_11_PhD_thesis.pdf.

Mossin, E., Pantoni, R., & Brandão, D. (2010). Students' evaluation based on fuzzy sets theory. In A.T. Azar (Ed.), *Fuzzy Systems*. Retrieved from

[http://www.intechopen.com/books/fuzzy-systems/students-evaluation-based-on-fuzzy-sets-theory.](http://www.intechopen.com/books/fuzzy-systems/students-evaluation-based-on-fuzzy-sets-theory)

Musavian, S. & Ahmadi, A. (2013). Assessment of Laboratory Courses Using Fuzzy reasoning. *International Research Journal of Applied and Basic Sciences* 6(5), 533-537.

Nasira, G., Kumar, S., & Kiruba, M. (2008). A comparative study of fuzzy logic with artificial neural networks algorithms in clustering. *Journal of Computer Application* 1(4), 6-8.

National Council on Teacher Quality (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies*. Retrieved from http://nctq.org/dmsView/State_of_the_Stes_Teacher_Evaluation_and_Effectiveness_Policies_NCTQ_Report.

National Council on Teacher Quality (2013). *State teacher policy yearbook: National summary*. Washington, DC: National Council on Teaching Quality. Retrieved from http://www.nctq.org/dmsView/2013_State_Teacher_Policy_Yearbook_Georgia_NCTQ_Report.

New Jersey Department of Education. (2014). *AchieveNJ: Teacher evaluation scoring guide*. Retrieved from <http://www.nj.gov/education/AchieveNJ/resources/TeacherEvaluationScoringGuide.pdf>.

Neogi, A., Mondal, A., & Mandal, S. (2011). A cascaded fuzzy inference system for university non-teaching staff performance appraisal. *Journal of Information Processing Systems*, 7(4), 595- 612.

New Teacher Project. (2010). *Teacher evaluation 2.0*. The New Teacher Project. Retrieved from <http://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf>.

New Teacher Project. (2013). *Fixing classroom observations: How Common Core will change the way we look at teaching*. Retrieved from <http://tntp.org/publications/view/fixing-classroom-observations-how-common-core-will-change-teaching>.

Nykänen, O. (2006). Inducing fuzzy models for student classification. *Journal of Educational Technology & Society*, 9(2), 223-234.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education*, 79 (4), 4-32.

Pappis, C. & Siettos, C. (2005). Fuzzy reasoning. In Burke & Kendall (Eds.), *Search Methodologies* (p. 437-474). US: Springer.

Partee, G. L. (2012). *Using multiple evaluation measures to improve teacher effectiveness: State strategies from round 2 of No Child Left Behind Act waivers*. Center for American Progress. Retrieved from <https://www.americanprogress.org/issues/education/report/2012/12/18/48368/using-multiple-evaluation-measures-to-improve-teacher-effectiveness/>.

Pavani, S.P., Gangadhar, P.V., & Gulhare, K.K. (2012). Evaluation of teacher's performance using fuzzy logic techniques. *International Journal of Computer Trends and Technology*, 2, 200-205.

Peterson, K.D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal*, 24(2), 311-317.

Peterson, K.D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. Corwin Press.

Peterson, K. (2004). Research on school teacher evaluation. *NASSP Bulletin* 88(639), 60-79.

Pianta, R.C., La Paro, K.M., & Hamre, B.K. (2006). *Classroom assessment scoring system: Observation protocol manual*. Baltimore, MD: Brookes.

Prince, C., Koppich, J., Azar, T., Bhatt, M., & Witham, P. (2013). *How well do principals' evaluations of teachers predict student achievement outcomes?* Center for Educator Compensation Reform. Retrieved from www.cecr.ed.gov/guides/.../Research%20Synthesis_Q%20D20.pdf

Purnama-Dewi, Oka-Sudana, & Darma-Putra (2012). Comparing scoring and fuzzy logic method for Teacher Certification DSS in Indonesia. *International Journal of Computer Science Issues (IJCSI)*, 9(6), 309-321.

Ramli, N. (2009). A centroid-based performance evaluation using aggregated fuzzy numbers, *Applied Mathematical Sciences*, 3(48), 2369-2381.

Reiss A., Hennessey, J.G., Rubin, M., Beach, L., Abrams, M.T., & Warsofsky, I.S. (1998). Reliability and validity of an algorithm for fuzzy tissue segmentation of MRI. *Journal of Computer Assisted Tomography* 22, 471-479.

Ribeiro, P. (n.d.). *Getting started with fuzzy logic*. Retrieved from <https://www.calvin.edu/~pribiero/othrlnks/Fuzzy/apps.htm>

Rickman, D. & Olivarez, E. (2014). Georgia Department of Education (2014). *Race to the Top: Georgia's vision for education excellence*. Retrieved from

<http://www.gadoe.org/Race-to-theTop/Documents/Race%20to%20the%20Top%20Four%20Year%20Report%20by%20GPEE.pdf>

RMC Research Corporation. (2013). *Evaluation report: Governor's Virginia performance pay incentives initiative pilot*. Richmond, VA: Virginia Department of Education.

Rockoff, J.E. & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *The American Economic Review*, 2, 261-271.
doi:10.1257/aer.100.2.261.

Rockoff, J.E., Staiger, D.O., Kane, T.J., & Taylor, E.S. (2011). Information and employee evaluation: Evidence from a randomized intervention in public schools. Working Paper 16240. American Economic Review. Retrieved from <http://www.nber.org/papers/w16240>.

Saleh, I. & Kim, S. (2009). A fuzzy system for evaluating students' learning achievement *Expert Systems with Applications*, 36 (2009), 6236–6243.

Salleh, M., Amin, A., Muda, S., Sofian, M., & Halim, A. (2013). Fairness of performance appraisal and organizational commitment. *Asian Social Science*, 9(2), 121-128.

Sanders, W.L. & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. University of Tennessee Value-Added Research and Assessment Center. Retrieved from http://www.cgp.upenn.edu/pdf/Sanders_Rivers-TVASS_teacher%20effects.pdf.

Schochet, P.Z. & Chiang, H.S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. Retrieved from <http://ies.ed.gov/ncee/pubs/20104004/>.

Semerçi, Ç. (2004). The influence of fuzzy logic theory on students' achievement. *Turkish Online Journal of Educational Technology*, 3(2), 56-61.

Shrout, P. & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.

Singh, R. & Pratap, V. (2011). Modeling academic performance evaluation using soft computing techniques: A fuzzy logic approach. *International Journal on Computer Science and Engineering*, 2, 676-686.

Singh, H., Gupta, M. Meitzler, T., Hou, Z., Garg, K. Solo, A., & Zadeh, L. (2013). Real-life applications of fuzzy logic. *Advances in Fuzzy Systems, 2013*, 1-3. doi:10.1155/2013/581879.

Sripa, R.R. & Suksawat, B.B. (2010). Propose of fuzzy logic-based students' learning assessment. *2010 International Conference on Control Automation & Systems (ICCAS)*, 414-424.

Srivastava, A., Rastogi, A., Srivastva, V.K., Saxena, K., & Arora, S. (2010). Analyzing motivation of private engineering college students: A fuzzy logic approach (A case study of private engineering college). *International Journal on Computer Science and Engineering*, 2(4), 1467-1476.

Stronge, J.H. & Tonneson, V.C. (2011). *CLASS Keys Teacher Evaluation System recommendations for improvement*. Atlanta, GA: Georgia Department of Education. Retrieved from <https://www.strongeandassociates.com/articles.html>.

Stronge, J.H., Ward, T.J., & Grant, L.W. (2011). What makes good teachers good? A cross- case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, 62(4), 339-355.

Stronge, J.H., Ward, T.J., Tucker, P.D., & Hindman, J.L. (2008). What is the relationship between teacher quality and student achievement? An exploratory study. *Journal of Personnel Evaluation in Education*, 20, 165-184.

Stronge, J.H., Xu, X., Leeper, L. & Tonneson, V.C. (2013). *Strong teacher evaluation system: A validation report*. Retrieved from http://www.cesa6.org/effectiveness_project/Validation-Report-of-Stronge-Evaluation-System.pdf.

Szmidt, E., Kacprzyk, J., & Bujnowski, P. (2012). On an enhanced method for a more meaningful Pearson's correlation coefficient between intuitionistic fuzzy sets. *Artificial Intelligence & Soft Computing*, 334-354. doi:10.1007/978-3-642-29347-4_39.

Taylor, E.S. & Tyler, J.H. (2012). The effect of evaluation on teacher performance: Evidence from longitudinal student achievement data of mid-career teachers. Working Paper w16877. National Bureau of Economic Research.

Taylan, O. & Karagözoğlu, B. (2009). An adaptive neuro-fuzzy model for prediction of student's academic performance. *Computers & Industrial Engineering*, 57(3), 732-741. doi:10.1016/j.cie.2009.01.019.

The Whitehouse. (2014). *Race to the Top*. Retrieved from <http://www.whitehouse.gov/issues/education/k-12/race-to-the-top>.

Toch, T. & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education. Education sector reports*. Washington, DC: Education Sector. Retrieved from http://www.educationsector.org/usr_doc/RushToJudgment_ES_Jan08.pdf

Trstenjak, B. & Donko, D. (2013). Teacher quality evaluation in HEI using a fuzzy logic. Paper presented at Conference of Informatics and Management Sciences, March 25-29, 2013, 319-326.

Tyler, J.H. (2011). *Developing high quality evaluation systems for high school teachers*. Center for American Progress. Retrieved from <https://www.americanprogress.org/issues/education/report/2011/11/29/10614/designing-high-quality-evaluation-systems-for-high-school-teachers/>.

U.S. Department of Education (2009). *Race to the Top: Executive summary*. Washington, DC: U.S. Department of Education. Retrieved from <http://www2.ed.gov/programs/racetothetop/index.html>.

von Altrock, C. (1994). Fuzzy logic and neurofuzzy in appliances. In proceedings of the 1994 Embedded Systems Conferences, Santa Clara, CA, 1994, 1-10.

Voskoglou, M.V. (2013). Fuzzy logic as a tool for assessing students' knowledge and skills. *Education Sciences*, 3(2), 208-221. doi:10.3390/edusci3020208.

Wang, C. & Chen, S. (2008). Appraising the performance of high school teachers based on fuzzy number arithmetic operations. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 12(2008), 919-934.

Wang, X., Wong, K., & Wong, J.Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology*, 95(3), 546-561. doi:10.1037/a0018866.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teachers Project. Retrieved from <http://widgeteffect.org/downloads/TheWidgetEffect.pdf>.

Weiss, E. & Long, D. (2013). *Market-oriented education reforms' rhetoric trumps reality*. Retrieved from <http://www.epi.org/files/2013/bba-rhetorictrumps-reality.pdf>.

Weon, S. & Kim, J. (2001). Learning achievement evaluation strategy using fuzzy membership function. In Proceedings of the 31st ASEE/IEEE Frontiers in Education Conference, Reno, NV, 19-24.

Whitehurst, G., Chingos, M., & Lindquist, K. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Brookings Institution. Retrieved from <http://www.brookings.edu/research/reports/2014/05/13-teacher-evaluation-whitehurst-chingos>.

Wood, J., Tocci, C., Joe, J., Holzman, S., Cantrell, S., & Archer, J. (2014). *Building trust in observations: A blueprint for improving systems to support great teaching*. Measures of Effective Teaching (MET) Project. (2012). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from www.metproject.org/downloads/MET_Observation_Blueprint.pdf.

Wu, J., Tsai, H., Shih, M., & Fu, H. (2010). Government performance evaluation using a balanced scorecard with a fuzzy linguistic scale. *Service Industries Journal*, 30(3), 449-462. doi:10.1080/02642060802248017.

Yadav, R., Ahmed, P.P., Soni, A.K., & Pal, S. (2014). Academic performance evaluation using soft computing techniques. *Current Science (00113891)*, 106(11), 1505-1517.

Yadav, R. & Singh, V.P. (2011). Modeling academic performance evaluation using soft computing techniques: A fuzzy logic approach. *Internatiaonl Journal of Computer Science and Engineering* 3(2), 676-686.

Yates, D. (2009). *Using fuzzy logic to identify schools which may be misclassified by the No Child Left Behind Adequate Yearly Progress policy*. Available from ProQuest Dissertations and Theses database (UMI No. 3367537), (Doctoral dissertation).

Yin, R. (1994). *Case study research: Design and methods*. Thousand Oaks, CA: Sage Publishing.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, 8, 338-353.

Zadeh, L.A. (1975). The concept of a linguistic variable and its application to appropriate reasoning. *Information Sciences*, 8, 43-80.

Zadeh, L.A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, 9(1), 149-184.

Zhao, J. & Bose, B.K. (2002). Evaluation of membership functions for fuzzy logic controlled induction motor drive. Paper presented at the 28th Annual Conference of the Industrial Electronics Society, 229-234. doi:10.1109/IECON.2002.118751.

Zubrzycki, J. (2012). Districts abandon grants targeting teacher quality. *Education Week* 32(1), 20-21.

APPENDIX A: A Survey of TAPS Standards Affecting Teacher's Performance in Secondary Education

Instructions: Teacher Assessment on Performance Standards (TAPS) is one component of the Teacher Keys Effectiveness System, which provides a qualitative, rubrics-based evaluation method to measure teacher performance related to quality performance standards used in this survey.

As an Expert in this area and using the following definitions of the Teacher Assessment of Performance Standards (TAPS) you are requested to rate to what extent, in your opinion, each standard is critical to teacher performance with the scale 1-10 as defined. Please mark your response in the area provided on the back of this paper. Your timely response is appreciated.

10	1	
Most Critical to teacher's performance		Least Critical teacher's performance
<p>1. Professional Knowledge: <i>The teacher demonstrates an understanding of the curriculum, subject content, pedagogical knowledge, and the needs of students by providing relevant learning experiences.</i></p>		
<p>2. Instructional Planning: <i>The teacher plans using, state and local school district curricula and standards, effective strategies, resources, and data to address the differentiated needs of all students.</i></p>		
<p>3. Instructional Strategies: <i>The teacher promotes student learning by using research-based instructional strategies relevant to the content to engage students in active learning and to facilitate the students' acquisition of key knowledge and skills.</i></p>		
<p>4. Differentiated Instruction: <i>The teacher challenges and supports each student's learning by providing appropriate content and developing skills which address individual learning differences.</i></p>		
<p>5. Assessment Strategies: <i>The teacher systematically chooses a variety of diagnostic, formative, and summative assessment strategies and instruments that are valid and appropriate for the content and student population.</i></p>		
<p>6. Assessment Uses: <i>The teacher systematically gathers, analyzes, and uses relevant data to measure student progress, to inform instructional content and delivery methods, and to provide timely and constructive feedback to both students and parents.</i></p>		
<p>7. Positive Learning Environment: <i>The teacher provides a well-managed, safe, and orderly environment that is conducive to learning and encourages respect for all.</i></p>		
<p>8. Academically Challenging Environment: <i>The teacher creates a student-centered, academic environment in which teaching and learning occur at high levels and students are self-directed learners.</i></p>		
<p>9. Professionalism: <i>The teacher exhibits a commitment to professional ethics and the school's mission, participates in professional growth opportunities to support student learning, and contributes to the profession.</i></p>		

Instructions: Teacher Assessment on Performance Standards (TAPS) is one component of the Teacher Keys Effectiveness System, which provides a qualitative, rubrics-based evaluation method to measure teacher performance related to quality performance standards used in this survey.

As an Expert in this area and using the following definitions of the Teacher Assessment of Performance Standards (TAPS) you are requested to rate to what extent, in your opinion, each standard is critical to teacher performance with the scale 1-10 as defined. Please mark your response in the area provided on the back of this paper. Your timely response is appreciated.

10. Communication: *The teacher communicates effectively with students, parents or guardians, district and school personnel, and other stakeholders in ways that enhance student learning.*

- _____ Professional Knowledge
- _____ Instructional Planning
- _____ Instructional Strategies
- _____ Differentiated Instruction
- _____ Assessment Strategies
- _____ Assessment Uses
- _____ Positive Learning Environment
- _____ Academically Challenging Environment
- _____ Professionalism
- _____ Communication

APPENDIX B: Definition and Description of Teacher Ratings

Category	Description	Definition
Exemplary	The teacher performing at this level maintains performance, accomplishments, and behaviors that continually and considerably surpass the established performance standard, and does so in a manner that exemplifies the school's mission and goals. This rating is reserved for performance that is truly exemplary and is demonstrated with significant student learning gains.	Exemplary performance: Continually meets the standards empowers students and exhibits continuous behaviors that have a strong positive impact on student learning and the school climate acquires and implements new knowledge and skills and continually seeks ways to serve as a role model to others
Proficient	The teacher meets the performance standard in a manner that is consistent with the school's mission and goals and has a positive impact on student learning gains.	Proficient performance: Consistently meets the standards engages students and exhibits consistent behaviors that have a positive impact on student learning and the school climate demonstrates willingness to learn and apply new skills
Needs Development	The teacher inconsistently performs at the established performance standard or in a manner that is inconsistent with the school's mission and goals and may result in below average student learning gains. The teacher may be starting to exhibit desirable traits related to the standard, but due to a variety of reasons, has not yet reached the full level of proficiency expected or the teacher's performance is lacking in a particular area.	Needs Development performance: Requires frequent support in meeting the standards results in less than expected quality of student learning needs guidance in identifying and planning the teacher's professional growth
Ineffective	The teacher continually performs below the established performance standard or in a manner that is inconsistent with the school's mission and goals and results in minimal student learning gains.	Ineffective performance: Does not meet the standards results in minimal student learning may contribute to a recommendation for the employee not being considered for continued employment

Reprinted from Georgia Department of Education (2013). Copyright 2013 by Georgia Department of Education. Reprinted with permission.